Univerza v Ljubljani
Fakulteta za računalništvo in informatiko

Miha Svetelšek

Napovedovanje fenotipa iz podatkov o genotipu posameznikov in celotnih generacij

diplomsko delo

univerzitetni interdisciplinarni program
računalništva in matematike

Mentor: doc. dr. Tomaž Curk

Ljubljana 2014

Besedilo je oblikovano z urejevalnikom besedil LaTeX.

Fakulteta za računalništvo in informatiko izdaja naslednjo nalogo:

Tematika naloge:

V diplomski nalogi preučite možnost modeliranja povezave med genotipom in fenotipom. Uporabite podatke o osemindvajsetih vzorcih posameznikov in dveh populacij kvasovke *S. cerevisiae*. Določite najmanjši nabor mutacij posameznih nukleotidov in genov, na podlagi katerih je možno zgraditi dober napovedni model za fenotip. Preverite ali uporaba predznanja o funkcijah genov pripomore k izgradnji boljših napovednih modelov. Empirično določite minimalno število vzorcev posameznikov in populacij, ki so potrebni za izgradnjo dobrega napovednega modela. Za modeliranje uporabite linearno in logistično regresijo ter poročajte o napaki napovedi.

Izjava o avtorstvu diplomskega dela

Spodaj podpisani Miha Svetelšek, z vpisno številko **63070092**, sem avtor diplomskega dela z naslovom:

Napovedovanje fenotipa iz podatkov o genotipu posameznikov in celotnih generacij

S svojim podpisom zagotavljam, da:

- sem diplomsko delo izdelal samostojno pod mentorstvom doc. dr. Tomaža Curka,
- so elektronska oblika diplomskega dela, naslov (slov., angl.), povzetek (slov., angl.) ter ključne besede (slov., angl.) identični s tiskano obliko diplomskega dela
- soglašam z javno objavo elektronske oblike diplomskega dela v zbirki "Dela FRI".

V Ljubljani, dne 23. junija 2014 Podpis avtorja:

Na tem mestu bi rad izkoristil priložnost in se zahvalil vsem, ki me trpite, ste z mano potrpežljivi, me imate radi in mi pomagate pri življenjskih odločitvah.

Posebna zahvala gre mentorju, doc. dr. Tomažu Curku, ki mi je velikokrat svetoval in pomagal.

Družini.

Kazalo

Seznam uporabljenih kratic

kratica	angleško	slovensko
GO	gene ontology	ontologija funkcijskih pripisov genov
KEGG	Kyoto encyclopedia of genes and genomes	kjotska enciklopedija genov in genomov
IP	inferior parent	manjvredni starš
SP	superior parent	večvredni starš
SNV	single nucletoide variant	varianta posameznega nukleotida
MAE	mean absolute error	povprečna absolutna napaka
MSE	mean squared error	povprečna kvadratna napaka
ESS	error sum of squares	napaka vsot kvadratov
SSE	sum of squares error	vsota kvadratov ostankov
FDR	false discovery rate	delež napačno pozitivnih zadetkov

Povzetek

V diplomski nalogi smo modelirali povezavo med genotipom in fenotipom tridesetih vzorcev kvasovke *S. cerevisiae*. Na podlagi podatkov in predznanja smo določili mutacije posameznih nukleotidov in z njimi povezane gene, s katerimi je možno zgraditi dober model za napovedovanje fenotipa. Poleg določanja pomembnih mest v genomu (SNV-jev) nam zgrajeni model omogoča tudi določevanje pomembnih genotipov oziroma starševskega izvora, ki je povezan z opazovanim fenotipom. Vrednotenje modelov pokaže, da lahko z linearno regresijo zanesljivo napovedujemo fenotip. Fenotip relativno dobro napoveduje tudi model, ki je zgrajen le na podlagi podatkov o dveh izvornih starših in začetne populacije. Empirično smo določili povezavo med številom vzorcev, ki jih uporabimo za izgradnjo napovednih modelov, in napovedno napako modelov.

Ključne besede: bioinformatika, genotip, fenotip, posameznik, populacija, linearna regresija, logistična regresija.

Abstract

We have modeled the relationship between genotype and phenotype using data on thirty yeast *S. cerevisiae* samples. Using prior knowledge, we have determined mutations of individual nucleotides and related genes with which it is possible to build a good prediction model for the phenotype. The constructed models allow us to determine the location of important mutations in the genome (SNVs) and to determine significant genotypes or parental origin, which is connected to the observed phenotype. Evaluation of these models shows that the phenotype can be predicted very reliably with linear regression. The phenotype can be predicted relatively well from data on two starting parents and the first generation of segregants. We also show the relation between the number of samples used to build a predictive model and its predictive error.

Keywords: bioinformatics, genotype, phenotype, individual segregant, pool of segregants, linear regression, logistic regression.

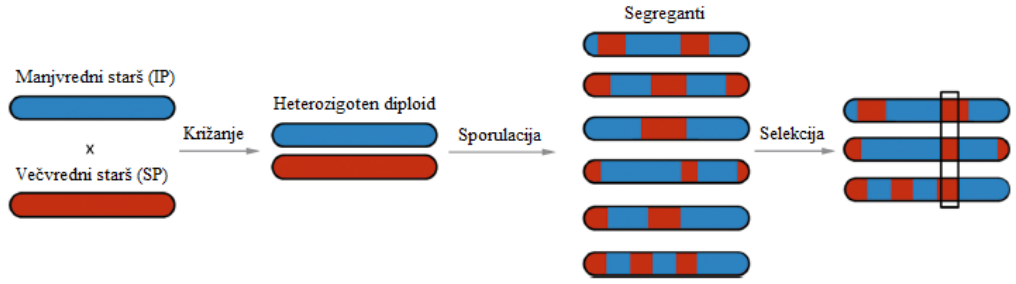
Poglavje 1

Uvod

Osnova diplomske naloge so eksperimentalni podatki o križanju različnih posameznikov kvasovke *S. cerevisiae*. Glavni cilj eksperimenta je bil izbrati potomce, ki so zelo odporni na določeno kemikalijo. Začetna posameznika populacije sta bila manjvredni starš (ang. IP - inferior parent), ki je slabo odporen na kemikalijo, in večvredni starš (ang. SP - superior parent), ki je dobro odporen na kemikalijo. Prva generacija potomcev je bila rezultat križanja klonov teh dveh posameznikov na različne načine (poimenovana $F1\_pool$ oziroma celotna populacija po prvem križanju). Vsaka naslednja generacija potomcev je bila dobljena tako, da so izbrali posameznike iz trenutne generacije, ki so bili najbolj odporni na kemikalijo, in jih medsebojno križali. Eksperiment so zaključili po sedmih generacijah križanja (poimenovana $F7\_pool$ oziroma celotna populacija po sedmih križanjih) ter v končni generaciji izbrali 26 najboljših posameznikov (urejeni po padajoči odpornosti na kemikalijo in poimenovani $F7JP\_01$ do $F7JP\_26$). Primer postopka križanja in genomov posameznikov zadnje generacije so prikazani na sliki 1.1. Nekateri deli genoma so torej podedovani od večvrednega starša, drugi pa od manjvrednega starša.

Začetne razlike med staršema so relativno majhne. Celoten genom kvasovke obsega 12,5 M baznih parov, od tega sta se začetna starša razlikovala na približno 70 K mestih oziroma le v $0,56\%$ celotnega genoma. Tovrstnim

mestom, kjer v neki populaciji obstaja variacija v baznem paru, pravimo različica posameznega nukleotida (ang. SNV - single nucleotide variant).



Slika 1.1: Genska rekombinacija, povzeto po Sliki 1 v [5]

1.1 Cilji diplomske naloge

V diplomski nalogi odgovarjamo na naslednja vprašanja:

1. Kako določiti najmanjši nabor genov oz. SNV-jev, ki bo čim boljše napovedoval fenotip vzorcev?

2. Kateri geni oz. SNV-ji so najbolj povezani s fenotipom vzorcev in od katerega od dveh začetnih staršev (večvredni ali manjvredni) morajo biti podedovani, da je fenotip posameznika tudi dober?

3. Na kakšen način moramo izbirati posameznike in populacije ter koliko jih potrebujemo za izgradnjo uspešnega napovednega modela, s katerim lahko dobro rangiramo ostale vzorce?

4. Kako dobro rangiramo posameznike iz končne populacije ($F7JP\_01$ do $F7JP\_26$), če napovedni model zgradimo le na podlagi podatkov o začetnih starših (IP in SP) in začetne populacije ($F1\_pool$)?

1.2 Metode dela

Pri odgovarjanju na zgornja vprašanja smo uporabili različne tehnike **strojnega učenja** (ang. machine learning) in **podatkovnega rudarjenja** (ang. data mining) [12, 11, 16, 19].

Surovi podatki navajajo genotip (bazni par) na vseh 70 K SNV-jih v genomu za oba začetna starša (IP, SP), populacijo po prvem križanju ($F1\_pool$) in po sedmem križanju ($F7\_pool$) ter za vseh 26 najboljših posameznikov, ki so bili izbrani iz zadnje populacije $F7\_pool$ ($F7JP\_01$ do $F7JP\_26$).

Osnovni podatki vključujejo tudi podatek o fenotipu vzorcev (posameznikov oziroma populacij). Nižja vrednost fenotipa predstavlja boljšo odpornost vzorca na kemikalijo. Ker je atributov (SNV-jev je 70 K) bistveno več kot vzorcev (posameznikov ali populacij je le 30), imamo opravka z neuravnoteženi podatki, kar predstavlja velik izziv pri modeliranju in doseganju visoke točnosti napovedovanja ter rangiranja fenotipov. Zaradi tega smo se na začetku osredotočili na postopke za zmanjšanje števila atributov (SNV-jev).

Za zmanjšanje števila SNV-jev smo uporabili predznanje iz dveh podatkovnih zbirk: pripisi funkcij genov GO (ang. GO - Gene Ontology [2]) in enciklopedije genov ter metabolnih procesov KEGG [10]. Podatke o genih iz obeh baz smo zapisali v obliki tabel, kjer je za vsak gen zapisano, v katerih procesih sodeluje. Na osnovi teh podatkov smo določili skupine genov, ki sodelujejo pri čim večjem številu istih procesov.

Gradili smo napovedne modele in nato s prečnim preverjanjem (definicija 3.7 [17, 22, 12, 11]) ovrednotili bodisi natačnost klasifikacije bodisi natančnost rangiranja vzorcev. Uporabili smo dva napovedna modela: logistično (definicija 3.5 [12, 8]) ter linearno regresijo (definicija 3.5 [12, 11]). Slednjo smo uporabili samo za klasificiranje vzorcev, medtem ko smo prvo uporabili za rangiranje vzorcev. Uspešnost napovednih modelov, zgrajenih iz naborov genov oz. SNV-jev, ki smo jih dobili s pomočjo predznanja, smo ocenili z dvema merama napak: povprečno absolutno napako (definicija 3.9

[20, 9]) in povprečno kvadratno napako (definicija 3.8 [23, 9]).

Zaradi preprostosti, dobrih orodij in raznovrstnih knjižnic smo kodo v celoti napisali v programskem jeziku Python. Za potrebe strojnega učenja in rudarjenja smo uporabili implementacije iz programske knjižnice *scikit-learn* [15].

Poglavje 2

Opis in procesiranje podatkov

Podatki so pridobljeni na podlagi sekvenciranja nove generacije, s katerim lahko na cenovno relativno ugoden način pridobimo zaporedje genoma v nekem vzorcu celic, kar lahko uporabimo za preučevanje povezave med genotipom in fenotipom [18, 4]. Ker so podatki in rezultati zaupni, v diplomski nalogi navajamo le tiste podrobnosti, ki so potrebne za razumevanje problema.

V naslednjem poglavju smo opisali metode, s katerimi smo odstranili manj pomembne SNV-je in tako dobili končno množico podatkov, na kateri smo gradili in vrednotili različne napovedne modele ter ovrednostili vpliv uporabe predznanja o funkcijah genov za boljše napovedovanje fenotipa.

2.1 Opis podatkov

Podatki vsebujejo genome in fenotipe 30 vzorcev: *manjvrednega starša* (IP - inferior parent), *večvrednega starša* (SP - superior parent), *dveh populacij* ($F1\_pool$ in $F7\_pool$) in *26 posameznikov* iz $F7\_pool$ ($F7JP\_01$ do $F7JP\_26$). Vzorec $F1\_pool$ predstavlja povprečje celotne prve generacije.

Ko smo pregledali populacijo $F7\_pool$, smo ugotovili, da je bila ustvarjena umetno. Določena je bila kot povprečje najboljših 26 posameznikov iz sedme generacije ($F7JP\_01$ - $F7JP\_26$). Ker ne nosi nobene dodatne informacije,

smo jo **odstranili** iz nadaljnje obravnave. Ostane nam torej 29 vzorcev in za vsakega izmed njih imamo podan celoten *genom*, ki je sestavljen iz 17 *kromosomov*. Ti so zgrajeni iz različnih *genov*, ki pa so sestavljeni iz več *SNV-jev* ('single nucleotide variant'). SNV-ji so torej osnovni opisni gradniki genoma.

Genomi vzorcev predstavljajo **genotip** in so podani v dvodimenzionalni tabeli, kjer 70913 vrstic predstavlja SNV-je (single nucleotide variant), in 397 stolpcev predstavlja različne atribute SNV-jev. Izmed vseh teh stolpcev smo pri svojem delu uporabili:

1. *snv_id* unikatno identifikacijsko ime SNV-ja (od snv00000 do snv70912),
2. *chrome* kromosom, na katerem je SNV (kromosomov je 17, od chrI do chrXVI in še chrM),
3. *dist_next* razdaljo do naslednjega SNV-ja, izraženo v številu nukleotidov (razdalje so pozitivne, zadnji SNV na posameznem kromosomu nima določene vrednosti, ker mu ne sledi noben SNV),
4. *gene* gen, v katerem je SNV (**Medgenski** SNV-ji imajo to polje prazno),
5. *cov_sum_all_samples* vsoto pokritosti mesta preko vseh vzorcev.

Uporabili smo tudi stolpce tipa *sample_closest_parent_sig*, ki za vsak vzorec in SNV podajajo, kateremu od izvornih staršev je vzorec bolj podoben. Vrednosti v tem stolpcu so lahko IP (opazovani SNV v tem vzorcu je bolj podoben manjvrednemu staršu), SP (opazovani SNV v tem vzorcu je bolj podoben večvrednemu staršu) ali *undet* (za opazovani SNV se ne moramo odločiti, kateremu staršu je bolj podoben). Te vrednosti zaradi lažjega dela spremenimo v diskretne celoštevilske, in sicer:

$$\begin{cases} 1, & \text{če je v i-tem SNV-ju j-tega vzorca, vrednost SP,} \\ -1, & \text{če je v i-tem SNV-ju j-tega vzorca, vrednost IP,} \\ 0, & \text{sicer.} \end{cases}$$

Fenotip vzorca je določen s celoštevilsko vrednostjo (rangom) na intervalu $[1, 29]$. Bolj odporni vzorci imajo nižjo vrednost. Velja tudi, da nobena dva vzorca nimata istega fenotipa.

2.2 Procesiranje podatkov

Atributov (SNV-jev) v podatkih je bistveno več kot primerov (vzorcev). Pri tako neuravnoteženih podatkih se lahko zgodi, da so pomembni atributi ne pridejo do izraza, saj so skriti med kopico nepomembnih. Težavo v podatkih predstavljajo tudi nedoločena genotipizacija vzorcev na posameznih SNV-ji (vrednost genotipa *undet* oz. 0). Ta mesta ne nosijo informacije o dedovanju oziroma o izvornem staršu, od kogar je bil del genoma podedovan, in so zato neuporabni za doseganje naših ciljev. Med procesiranjem podatkov smo želeli odstranili čimveč manj pomembnih SNV-jev in ohraniti čimveč pomembnih ter se znebiti čim večjega števila nedoločenih vrednosti.

Koraki procesiranja napisani v naslednjih razdelkih so opisani v enakem vrstnem redu, kot smo jih uporabili. Z uporabo drugačnega vrstnega reda dobimo slabše rezultate.

2.2.1 Glajenje podatkov

Genomi posameznikov, recimo manjvrednega in večvrednega starša, ne vsebujejo nedoločenih vrednosti. Populacija $F1\_pool$ pa jih vsebuje, ker predstavlja povprečje populacije prve generacije. Problematične nedoločene vrednosti so torej le v genomih posameznikov, pridobljenih iz populacije $F7\_pool$. Takih vrednosti je kar 295670 (približno 16% vseh primerov).

Znova poglejmo sliko 1.1. Opazimo, da so daljši, neprekinjeni deli kromosoma potomcev podobni enemu staršu, drugi pa drugemu staršu. To je rezultat procesa **genske rekombinacije**, kjer dedovanje ne poteka tako, da se deduje vsako mesto (SNV) posebej. Nasprotno, od enega starša potomec navadno podeduje daljše neprekinjeno področje naenkrat, ki lahko vsebuje tudi več genov. Poznavanje principov delovanja genske rekombinacije lahko

uporabimo za t. i. *glajenje podatkov*. Z glajenjem nedoločene vrednosti (*undet* oz. 0) v genomu zamenjamo z določenimi. Postopek, ki smo ga za to uporabili, je naslednji:

1. Vsak kromosom v vsakem vzorcu pregledamo in iščemo zaporedja SNV-jev oblike:

   IP-*undet*-...-*undet*-IP niz SNV-jev, katerih vrednost je nedoločena, je obdan z dvema SNV-jema, ki imata vrednost enako IP (podedovana od manjvrednega starša),

   SP-*undet*-...-*undet*-SP niz SNV-jev, katerih vrednost je nedoločena, je obdan z dvema SNV-jema, ki imata vrednost enako SP (podedovana od večvrednemega starša).

2. Empirično določimo zgornjo mejo (v nukleotidih) za dovoljeno oddaljenost med začetkom in koncem takšnih zaporedij.

3. Za vsako zaporedje torej izračunamo njegovo dolžino in jo primerjamo z zgornjo mejo za dovoljeno oddaljenost zaporedij. Podatke o oddaljenosti zaporednih SNV-jev dobimo v stolpcu **dist_next** v datoteki podatkov.

4. Vsa zaporedja, ki so krajša od zgornje meje dovoljene oddaljenosti, obravnavamo kot dele kromosoma, ki so podedovani od istega starša. Vsem nedoločenim SNV-jem znotraj takšnih zaporedij spremenimo vrednost. Tako zaporedja IP-*undet*-...-*undet*-IP postanejo zaporedja IP-IP-...-IP-IP in zaporedja SP-*undet*-...-*undet*-SP postanejo zapredja SP-SP-...-SP-SP.

Empirično smo zgornjo mejo oddaljenosti postavili na **500 nukleotidov**. Za takšno vrednost zgornje meje obstaja 105349 primerov, kjer nedoločene vrednosti spremenimo v določene. Delež nedoločenih vrednosti smo tako zmanjšali na 10.32%.

2.2.2 Odstranjevanje neinformativnih SNV-jev

V podatkih smo poiskali SNV-je, ki se v sedmih generacijah dedovanja **niso spremenili**. Torej tiste, ki imajo pri populaciji $F1\_pool$ enako vrednost genotipa kot pri vseh posameznikih iz populacije $F7\_pool$ (slika 2.1). SNV-ji s statičnim, nespreminjajočim genotipom so zelo verjetno nepovezani z opazovanim fenotipom in jih je zato smiselno odstraniti. To smo naredili z naslednjim postopkom:

1. Za vsak SNV preverimo genotip vseh 26 posameznikov iz populacije $F7\_pool$ ter genotip v populaciji $F1\_pool$.

2. Če so vse vrednosti genotipov enake, SNV odstranimo iz nadaljnje obravnave.

snv_id	F1_pool	F7JP_01	F7JP_02	F7JP_03	...	F7JP_24	F7JP_25	F7JP_26
...	...	...	...	...	...	...	...	...
snv17709	SP	SP	SP	SP	...	SP	SP	SP
...	...	...	...	...	...	...	...	...
snv54196	IP	IP	IP	IP	...	IP	IP	IP
...	...	...	...	...	...	...	...	...

Slika 2.1: Primeri neinformativnih SNV-jev v podatkih.

SNV-jev, ki se v sedmih generacijah dedovanja niso spremenili, je zelo veliko, saj po končanem zgornjem postopku ostane le še 41187 SNV-jev (58% začetnega nabora).

2.2.3 Združevanje podobnih zaporednih SNV-jev

Zaporedni SNV-ji, ki pripadajo istemu genu, so lahko zelo podobni ali celo enaki. V takih primerih je dovolj, če obdržimo le en SNV, saj nam ostali ne prinašajo nobene nove informacije. Postopek, ki smo ga uporabili za detekcijo in odstranjevanje podobnih, zaporednih SNV-jev je naslednji:

1. Empirično določimo spodnjo mejo za podobnost.

2. SNV-je, ki pripadajo istemu genu, pregledujemo zaporedoma in primerjamo i-ti SNV z (i+1)-im. Če sta si SNV-ja podobna bolj, kot je vrednost spodnje meje za podobnost, potem na seznam za odstranjevanje dodamo:

   - SNV z višjim indeksom (i+1) in
   - SNV z največjo pokritostjo preko vseh vzorcev (vrednostjo atributa *cov_sum_all_samples*).

   Drugi način se je izkazal za boljšega, saj odstranjuje SNV-je s slabšo pokritostjo.

3. Odstranimo vse SNV-je, ki so na seznamu za odstranjevanje.

Ker je vzorcev le 29, smo se v diplomski nalogi odločili, da bomo dva zaporedna SNV-ja združili le, če bosta popolnoma enaka (spodnja meja podobnosti je torej 29). Razlog za to je, da na tak način zagotovo ne izgubimo pomembnih SNV-jev. Tako nam po uporabi te metode ostane le še 37946 SNV-jev (53% začetnega nabora).

2.2.4 Odstranjevanje slabo pokritih SNV-jev

Pridobivanje podatkov o genomu vzorcev ni popolnoma zanesljiv proces. Pri sekvenciranju genomov vzorcev obstajajo tudi takšni SNV-ji, za katere nismo ravno prepričani, da smo jih prav odčitali. Vsoto pokritosti posameznega SNV-ja v vseh vzorcih (stolpec *cov_sum_all_samples*) lahko uporabimo kot mero zaupanja v pravilnost posameznega SNV-ja. Pokritost posameznega vzorca pove, koliko odčitkov, ki smo jih dobili pri sekvenciranju genoma, pokrije isti nukleotid.

Visoka vrednost vsote pokritosti vseh vzorcev v SNV-ju pomeni, da je veliko odčitkov pokrilo isti nukleotid oziroma SNV ter je zato bolj zaupanja

vreden kot neko drug nukleotid oziroma SNV, kjer je ta vsota nizka. Postopek je naslednji:

1. Empirično določimo spodnjo mejo za vsoto pokritosti vseh vzorcev.
2. Preverimo vsak SNV in na seznam za odstranjevanje dodamo tiste, katerih vsota pokritosti vseh vzorcev je manjša od spodnje meje.
3. Odstranimo vse SNV-je, ki so na seznamu za odstranjevanje.

Spodnjo mejo vsote pokritosti vseh vzorcev smo postavili na 70% deleža povprečne vsote pokritosti vseh SNV-jev. Število SNV-jev se je tako zmanjšalo na 29905 SNV-jev (42% začetnega nabora).
Opomba: *Pri pregledovanju podatkov smo opazili, da se povprečna vsota pokritosti vzorcev v različnih kromosomih precej spreminja (povprečna vsota pokritosti v* $chrVI = 1986.48$*,v* $chrM = 50209.9$*). Zato smo poskušali tudi tako, da smo za vsak kromosom posebej empirično določili spodnjo mejo za vsoto pokritosti. Kljub temu, da se nam je ta ideja zdela bolj smiselna, se tako dobljeni nabor SNV-jev ni bistveno razlikoval po napovedni vrednosti zgrajenih modelov.*

2.3 Iskanje osamelcev

V velikih zbirkah podatkov pogosto obstajajo primeri, ki so precej drugačni od vseh ostalih. Takšnim primerom rečemo **osamelci** (definicija 2.1) [24, 14]. Če klasificiramo ali rangiramo vzorce, ne da bi odstranili take primere, potem so naši rezultati manj točni. Zaradi tega moramo osamelce pred modeliranjem odstraniti.

Za odkrivanje osamelcev obstaja več orodij. Uporabili smo **Z-test** (definicija 2.2) [27] oziroma **Z-vrednost**, ki za posamezni primer pove, za koliko je v povprečju oddaljen od vseh ostalih.

Definicija 2.1 (Osamelec (outlier) [24])**.** *Osamelec definiramo kot opazovani primer v podatkih, ki je zelo različen od ostalih primerov.*

Definicija 2.2 (Z-vrednost [27])**.** *Z-vrednost je statistični kazalec položaja posamezne statistične enote v populaciji glede na aritmetično sredino. Izračuna se kot:*

$$z = \frac{x - \mu}{\sigma};$$

kjer je x trenutna vrednost, μ ***povprečna vrednost*** *(definicija 2.3) in σ* ***standardna deviacija*** *(definicija 2.4).*

Definicija 2.3 (Povprečna vrednost μ)**.** *Povprečna vrednost μ je definirana kot:*

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i;$$

kjer je N število primerov.

Definicija 2.4 (Standardna deviacija σ)**.** *Če je N število primerov, je standardna deviacija σ definirana kot:*

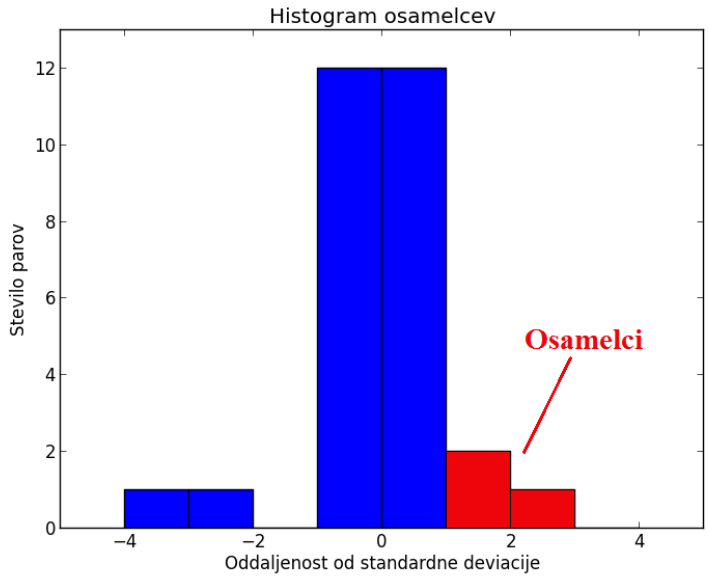
$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)}.$$

V podatkih lahko nastopata **dve različni populaciji** (eni posamezniki so bolj podobni manjvrednemu staršu, drugi pa bolj večvrednemu staršu). Zato smo Z-vrednosti računali glede na povprečne oddaljenosti vzorca do **treh najbližjih** vzorcev. Odstranili smo le tiste vzorce, ki so nadpovprečno oddaljeni do najbližjih vzorcev. To smo naredili z naslednjim postopkom:

1. Stolpce, ki jih definirajo vzorci, spremenimo v vektorje tako, da je vec_i vektor i-tega vzorca.

2. Dobljene vektorje razdelimo v pare tako: $par_{i,j} = (\vec{v_i}, \vec{v_j})$, kjer velja $1 \leq i < j \leq 29$ (vektorjev je 29, torej je parov $\frac{1}{2} * 29 * 28 = 406$).

3. Za vsak par vektorjev izračunamo njuno absolutno razliko $raz_{i,j} = |\vec{v_i} - \vec{v_j}|$.



Slika 2.2: Porazdelitev Z-vrednosti glede na oddaljenost vzorcev.

4. Nato za vsak vektor $\vec{v_i}$ najdemo tri vektorje $\vec{v_a}, \vec{v_b}$ in $\vec{v_c}$, za katere velja $raz_{i,a} \leq raz_{i,b} \leq raz_{i,c} \leq raz_{i,j}$ za katerkoli drug vektor $\vec{v_j}$.

5. Za vsak vektor $\vec{v_i}$ zdaj izračunamo njegovo povprečno razliko do njemu treh najbližjih vektorjev:

$$\overline{raz_i} = \frac{raz_{i,a} + raz_{i,b} + raz_{i,c}}{3}.$$

6. Izračunamo povprečno razliko med posameznim vektorjem in treh njemu najbližjih vektorjev:

$$\mu = \frac{1}{29} \sum_{i=1}^{29} \overline{raz_i}.$$

7. Izračunamo standardni odklon od povprečja σ:

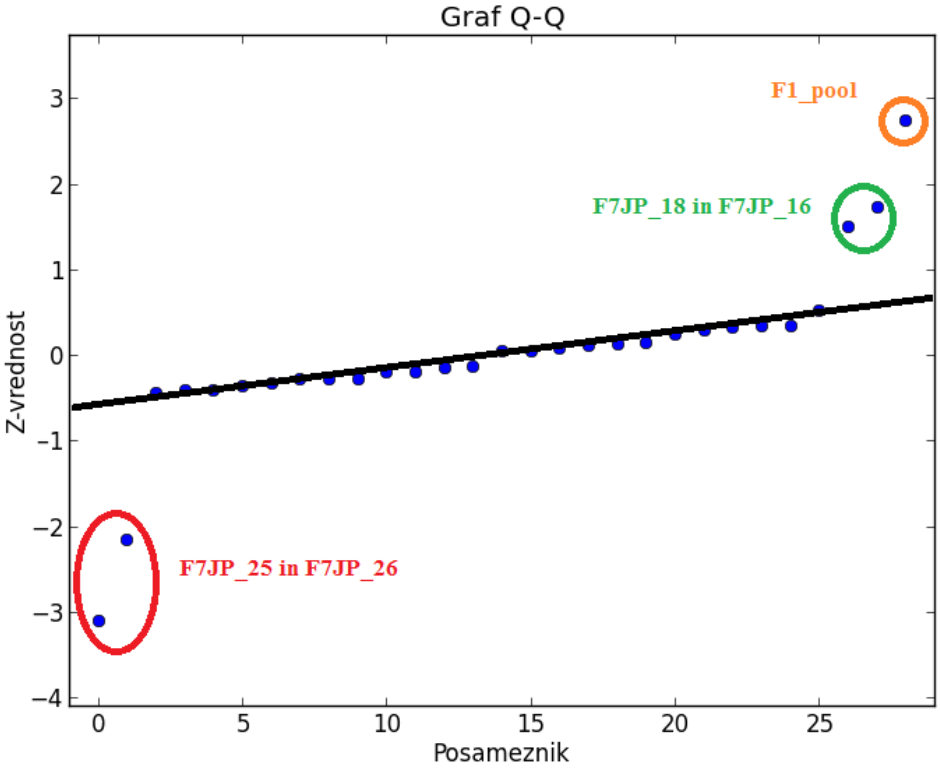
$$\sigma = \frac{1}{29} \sum_{i=1}^{29} (\overline{raz_i} - \mu).$$

8. Za vsak vektor $\vec{v_i}$ izračunamo Z-vrednost:

$$Zscore_i = \frac{\overline{raz_i} - \mu}{\sigma}.$$

9. Narišemo histogram dobljenih Z-vrednosti (slika 2.2).

10. Za lažje določanje osamelcev narišemo graf Q-Q (definicija 2.5) (slika 2.3).



Slika 2.3: Z-vrednosti vseh vzorcev.

Definicija 2.5 (Graf Q-Q [26])**.** *je oblika verjetnostnega grafa za primerjanje dveh porazdelitev tako, da v grafu primerjamo njune kvantile. Točka* (x, y) *na grafu pomeni, da kvantil druge verjetnostne porazdelitve (koordinata* y*) primerjamo s kvantilom prve verjetnostne porazdelitve (koordinata* x*). Če sta porazdelitvi podobni, bodo točke ležale na premici* $y = x$*. Če pa sta porazdelitvi linearno povezani, bodo točke ležale približno na isti premici.*

Na osnovi izračunanih Z-vrednosti vzorcev in grafičnih ponazoritev (slika 2.3 in 2.2) je možno informirano določiti osamelce:

Vzorec $F1\_pool$ je najbolj oddaljen od sebi najbližjih vzorcev, vendar ga ne odstranimo, ker predstavlja pomemben vzorec za odgovore na določena glavna vprašanja.

Vzorca $F7JP\_16$ in $F7JP\_18$ sta oba skoraj za dva standardna odklona (σ) bolj oddaljena do treh najbližjih vzorcev, kot so ostali vzorci do treh njim najbližjih vzorcev. Zatorej lahko trdimo, da sta osamelca in ju odstranimo.

Pri računanju Z-vrednosti smo ugotovili, da sta si vzorca $F7JP\_25$ in $F7JP\_26$ zelo podobna. Drugi vzorci so od njim najbližjih oddaljeni za najmanj 17000 mest, medtem ko sta $F7JP\_25$ in $F7JP\_26$ oddaljena le za 6769 mest. Ker sta si tako podobna, enega izmed njiju lahko odstranimo. Odločili smo se za vzorec $F7JP\_25$, saj je ta bolj podoben njemu najbližjim trem vzorcem.

Končni podatki tako izključujejo umetno sintetiziran vzorec $F7\_pool$, vzorec $F7JP\_25$ ter osamelca $F7JP\_16$ in $F7JP\_18$. Vrednosti fenotipov vzorcev so zato spremenjene (glej tabelo 2.1).

Vzorec	**Fenotip**	**Vzorec**	**Fenotip**
IP	26	F7JP_11	13
SP	10	F7JP_12	14
F1_pool	3	F7JP_13	15
F7JP_01	1	F7JP_14	16
F7JP_02	2	F7JP_15	17
F7JP_03	4	F7JP_17	18
F7JP_04	5	F7JP_19	19
F7JP_05	6	F7JP_20	20
F7JP_06	7	F7JP_21	21
F7JP_07	8	F7JP_22	22
F7JP_08	9	F7JP_23	23
F7JP_19	11	F7JP_24	24
F7JP_10	12	F7JP_26	25

Tabela 2.1: Fenotipi vzorcev po odstranitvi $F7\_pool$, vzorca $F7JP\_25$ ter osamelcev $F7JP\_16$ in $F7JP\_18$.

Poglavje 3

Metode

Število SNV-jev iz začetne zbirke podatkov smo v fazi procesiranja podatkov (prejšnje poglavje) zmanjšali na manj kot polovico (42%). Z uporabo predznanja smo želeli še dodatno skrčiti nabor SNV-jev, ki vstopajo v fazo gradnje napovednih modelov. V te nabore genov smo poskušali dodajati tudi medgenske SNV-je, za katere nimamo nobenega podanega predznanja.

V tem poglavju podamo definicijo obeh tipov napovednih modelov, ki smo jih uporabili za klasificiranje in rangiranje vzorcev. Opisana sta tudi potek gradnje napovednih modelov na podlagi predznanja in metoda za ocenjevanje uspešnosti napovednih modelov.

3.1 Priprava predznanja

Določiti smo želeli majhne nabore SNV-jev, ki bodo nadvse dobro napovedovali fenotipe vzorcev. Preveriti smo želeli idejo, da bi nabore SNV-jev določili na podlagi prisotnosti v genih, ki sodelujejo pri istih procesih. To znanje je dostopno na spletni strani *http://www.geneontology.org/*, kjer je za vsak gen podana informacija, v katerih bioloških procesih sodeluje. Z uporabo orodja *Orange* [3] smo podatke o genih pridobili, uredili in strnili v tabelo **GO**, ki obsega podatke o 5288 *genih* (v vrsticah) in o 39560 *bioloških procesih* ter drugih funkcijskih pripisih genov (v stolpcih).

Na podoben način smo s spletni strani *http://www.genome.jp/kegg/* pridobili tudi podatke o udeleženosti posameznih genov v metabolnih procesih in jih shranili v tabelo **KEGG**. Zgradba te tabele je podobna zgradbi tabele GO in obsega 1700 *genov* (v vrsticah) in 105 *metabolnih procesov* (v stolpcih).

Vrednosti v obeh tabelah so definirane na enak način: vrednost i-te vrstice j-tega stolpca X_{ij} v tabeli je definirana kot:

$$X_{ij} = \begin{cases} 1, & \text{če je i-ti gen pripisan j-ti funkciji oz. procesu,} \\ 0, & \text{sicer.} \end{cases}$$

Tabeli GO in KEGG (še posebej GO) sta sestavljeni iz velikega števila procesov in genov, kar je bilo treba zmanjšati, preden smo začeli z iskanjem naborov genov, ki sodelujejo pri istih procesih.

3.1.1 Procesiranje podatkov o pripisih funkcij genov

Izvorni tabeli GO in KEGG smo najprej spremenili tako, da sta vsebovali le tiste gene (vrstice), ki pripadajo SNV-jev, ki so se obdržali po začetnem procesiranju podatkov in filtriranju SNV-jev. Na ta način smo iz GO odstranili 376 genov (7% vseh genov v datoteki), iz KEGG pa 108 genov (6% vseh genov v datoteki).

Iz izvornih tabel GO in KEGG smo odstranili tudi funkcijske pripise (stolpce), ki so bodisi preveč specifični bodisi preveč splošni. Pri prvih sodeluje le par genov, pri drugih pa ogromno število genov. Taki procesi niso toliko informativni za razločevanje funkcije genov in le povečujejo časovno ter računsko zahtevnost. Odstranili smo jih po naslednjem postopku:

1. Najprej empirično določimo:
   - spodnjo mejo **low_bound** števila genov, ki morajo sodelovati pri nekem procesu, da ta ni preveč specifičen, in
   - zgornjo mejo **up_bound** števila genov, ki morajo sodelovati pri nekem procesu, da ta ni preveč splošen.

Tabela	up_bound	low_bound	% ostalih procesov
GO	25% * len(s_i)	20	29.6%
KEGG	25% * len(s_i)	5	90.6%

Tabela 3.1: Delež preostalih procesov za empirično določeni meji.

2. Za vsak proces p_i izračunamo, koliko genov pri njem sodeluje. Zaradi definicije vrednosti v tabeli (GO oz KEGG) to preprosto izračunamo kot vsoto stolpca:
$$sum(p_i) = \sum_{j=1}^{len(p_i)} .$$
3. Dobljene vsote primerjamo z mejama *low_bound* in *up_bound*. Iz tabele odstranimo stolpce, za katere ne velja:
$$low\_bound < sum(s_i) < up\_bound.$$

Empirično določeni meji in število procesov, ki nam ostane po uporabi zgornjega postopka, so podani v tabeli 3.1.

3.1.2 Razvrščanje v skupine

Gene v tabelah iz prejšnjega poglavja smo razvrstili v skupine podobnih genov, ki sodelujejo pri istih procesih, in sicer s pomočjo aglomerativnega tipa hierarhičnega razvrščanja v skupine (definicija 3.1 in 3.2) [21, 19]. Za združevalno metodo smo uporabili Wardovo metodo (definicija 3.3) [28, 13, 19]. Zaradi hitrosti in lažjega dela smo za razvrščanje v skupine uporabili implementacije iz razreda *sklearn.cluster* programski paketa *scikit-learn* [15].

Definicija 3.1 (Razvrščanje v skupine [21, 19])**.** *Je naloga, katere rešitev so skupine predmetov, za katere velja, da so si predmeti znotraj iste skupine med seboj bolj podobni, kot so podobni predmetom iz katerekoli druge skupine.*

Definicija 3.2 (Hierarhično razvrščanje v skupine [21, 19])**.** *Je metoda razvrščanja, katere cilj je hierarhična ureditev skupin. Poznamo dva tipa takega razvrščanja:*

Aglomerativni: to je pristop, kjer skupine gradimo iz dna proti vrhu. Na začetku je vsak primer v svoji skupini. V vsakem koraku nato združimo najbližji par skupin v eno skupino.

Razdvojevalni: to je ravno nasproten pristop od aglomerativnega. V prvem koraku so vsi primeri v eni skupini. V vsakem naslednjem koraku nato vsako skupino razdvojimo v dve skupini.

Definicija 3.3 (Wardova metoda [28, 13, 19])**.** *Je združevalna metoda, ki se uporablja pri aglomerativnem tipu hierarhičnega razvrščanja v skupine (definicija 3.2). Kot kriterij za razvrščanje pri tej metodi uporabimo napako vsot kvadratov ('Error Sum of Squares' ali **ESS** (definicija 3.4)). Postopek združevanje poteka tako:*

1. *V prvem koraku združevanja je $ESS = 0$, ker je vsak posameznik v svoji skupini.*
2. *V vsakem naslednjem koraku nato združimo tisti dve skupini, ki najmanj povečata vrednost napake ESS.*

Definicija 3.4 (Napaka vsot kvadratov ali ESS)**.** *Naj bo X_{ijk} vrednost spremenljivke k v koraku j, ki pripada skupini i, in naj bo $x_{i;k}$ povprečna vrednost skupine i za spremenljivko k. Potem ESS izračunamo kot:*

$$ESS = \sum_i \sum_j \sum_k |X_{ijk} - \overline{x}_{i;k}|^2.$$

Pomemben parameter razvrščanja v skupine je število pričakovanih skupin. Izbrali smo takšno število skupin, da je bila pričakovana velikost posamezne skupine 20 genov. Za razvrščanje genov na podlagi tabele GO smo izbrali pričakovano število skupin 240, za tabelo KEGG pa smo izbrali 80 skupin. Posamezne skupine smo tudi medsebojno združevali in preverili, ali modeli, zgrajeni na tako združenih skupinah genov, dosegajo boljšo napovedno natančnost.

3.1.3 Medgenski SNV-ji

SNV-je, ki jih ne moremo pripisati nobenemu genu, ker so preveč oddaljeni od znanih genov, imenujemo **medgenski SNV-ji**. V končnem naboru SNV-jev je bilo 5574 medgenskih (19% vseh SNV-jev). Ker teh SNV-jev ne moremo pripisati genom, ne moremo uporabiti predznanja v GO ali KEGG, da bi skrčili nabor.

Poskusili smo določiti način, kako bi skupinam genov, dobljenih iz predznanja GO in KEGG, dodali še medgenske SNV-je. Dodajanje posameznih medgenskih SNV-jev bi bilo preveč zamudno. Zato smo se odločili medgenske SNV-je razvrstiti v skupine:

1. Zaporedne medgenske SNV-je med dvema genoma združimo v en povprečen medgenski SNV $\overline{mg\_SNV_i}$. Iz k zaporednih medgenskih SNV-jev izračunamo medgenski SNV:

$$\overline{mg\_SNV_i} = \frac{1}{k} \sum_{j=1}^{k} SNV_j.$$

2. Empirično določimo število skupin, v katere želimo razvrstiti medgenske SNV-je.

3. Povprečne medgenske SNV-je $\overline{mg\_SNV_i}$ hierarhično razvrstimo v skupine z Wardovo metodo.

Medgenske SNV-je smo razvrstili v 100 skupin.

3.2 Regresijski modeli

Za napovedovanju vzorcev smo uporabili dva regresijska modela: **linearno** (definicija 3.5) [12, 11] in **logistično regresijo** (definicija 3.2) [12, 8]. Uporabili smo implementaciji iz razreda *sklearn.linear_model* programskega paketa **scikit-learn** [15].

Oba napovedna modela sta primera *posplošenih linearnih modelov* (ang. GLM - generalized linear models) [12], za katere velja splošna enačba:

$$\widehat{f(c)} = g(v_1, ...v_a) = w_0 + \sum_{i=1}^{a} w_i v_i = w^T v, \tag{3.1}$$

kjer je $f(c)$ poljubno izbrana funkcija (vezna funkcija) odvisne spremenljivke c in $v^T = \langle 1, v_1, ..., v_a \rangle$ vektor vrednosti atributov. Naloga je najti vektor w parametrov w_i, $i = 0, ..., a$, ki minimizirajo *vsoto kvadratov ostankov* (ang. SSE - sum of squarred errors):

$$SSE = \sum_{j=1}^{n} (cj - \widehat{c_j})^2 = \sum_{j=1}^{n} (c_j - w_0 - \sum_{i=1}^{a} w_i v_{i,j})^2, \tag{3.2}$$

kjer so c_j dejanski rezultati, $\widehat{c_j}$ pa napovedani rezultati za vsak $j = 1, ..., n$.

Definicija 3.5 (Linearna regresija [12, 11])**.** *Podanih imamo m primerov in n atributov. Z njimi zgradimo matriko A, ki je dimenzije $m \times n$, in rešujemo problem $Aw = b$, kjer je b matrika dimenzije $n \times 1$ z dejanskimi rezultati primerov. Linearna regresija uporabi enačbo 3.1, kjer za vezno funkcijo uporabimo: $f(c) = c$. Nato iščemo vektor $w = \langle w_0, w_1, ..., w_a \rangle$, da velja:*

$$Aw - b = 0.$$

Sistem $Aw = b$ v splošnem ni rešljiv, ko je število primerov večje od števila atributov. Takrat iščemo rešitev, ki bo minimizirala $Aw - b$. To naredimo tako, da rešujemo sistem:

$$(A^T A)w = (A^T b),$$

tako da iščemo minimalno vrednost

$$\min_w = ||Aw - b||_2^2.$$

Opomba: *Za linearno regresijo velja, da vse atribute obravnavamo kot zvezne, tudi če so ti diskretni.*

Definicija 3.6 (Logistična regresija [12, 8])**.** *Čeprav ime nakazuje, da se ta napovedni model uporablja za regresijo, pa je logistična regresija metoda*

za klasifikacijo. Uporablja se za napovedovanje verjetnosti rezultata glede na vrednosti atributov. Za vezno funkcijo se pri logistični regresiji uporablja funkcija ***logit****, ki je definirana tako:*

$$f(c) = log(\frac{c}{1-c}).$$

PRIMER: če imamo razreda C_1 in C_2, potem je funkcija logit razreda C_1:

$$f(c) = log(\frac{P(C_1)}{1-P(C_1)} = w^T v.$$

Če enačbo preuredimo, dobimo **sigmoidno** funkcijo:

$$y = P(C_1) = (1 + e^{-w^T v})^{-1}. \tag{3.3}$$

Da določimo parametre vektorja w, imamo podane učne primere $\Gamma = \{\langle t_l, d(l)\rangle\}$, $l = 1, .., n$, kjer je $d(l) = 1$, če je pravilni razred primera t_l enak C_1 in $d(l) = 0$, če je pravi razred primera $t(l)$ enak C_2. Predpostavimo, da $d(l)$ s podanim t_l sledi **Bernoullijevi** distribuciji z verjetnostjo $y(l) = P(C_1|t_l)$, če jo izračunamo z enačbo 3.3:

$$d(l)|t_l \sim Bernoulli(y(l)).$$

Verjetnost posameznega vzorca je nato definirana kot:

$$l(w|\Gamma) = \prod_l y(l)^{d(l)}(1 - y(l))^{1-d(l)}. \tag{3.4}$$

Če nato enačbo 3.4 logaritmiramo z negativnim predznakom, dobimo funkcijo napake E, v našem primeru je to funkcija križne entropije:

$$\begin{aligned} E(w|\Gamma) &= -log(l(w|\Gamma)) \\ &= -\sum_l d(l)log(y(l)) + (1 - d(l))log(1 - y(l)). \end{aligned}$$

Da zmanjšamo križno entropijo in tako maksimiziramo verjetnost primera, lahko uporabimo npr. gradientno metodo. Za sigmoidno funkcijo v enačbi 3.3 dobimo naslednjo enačbo za spreminjanje parametrov w_j:

$$\begin{aligned}\triangle w_j &= -\eta \frac{\partial E}{\partial w_j} \\ &= \eta \sum_l (d(l) - y(l)) v_{j,l}.\end{aligned}$$

Ko določimo parametre w_j, logistična regresija klasificira primer v tisti razred C_k, $k \in 1, 2$, ki ima napovedano največjo verjetnost:

$$\frac{P(C_k)}{1 - P(C_k)}.$$

V diplomski nalogi smo imeli več vrednosti rezultatov (klasificirali smo v več razredov), zato smo uporabili **multiplo** logistično regresijo. Vezna funkcija logit je pri multipli logističi regresiji definirana tako:

$$\begin{aligned}logit(p_i) = ln(\frac{p_i}{1 - p_i}) = \beta_0 + \beta_1 x_{1,i} + ... + \beta_k x_{k,i}, \\ \text{kjer je } p_i = (1 + e^{-(\beta_0 + \beta_1 x_{1,i} + ... + \beta_k x_{k,i})})^{-1}. \qquad (3.5)\end{aligned}$$

Do rešitve za vektor w pridemo tako, da analogno nadaljujemo postopek, ki smo ga opisali za primer, ko klasificiramo v dva razreda.
V diplomski nalogi smo linearno regresijo uporabili le za klasificiranje vzorcev, medtem ko smo logistično regresijo uporabili tudi za njihovo rangiranje.

3.3 Gradnja napovednih modelov

Definicija 3.7 (K-kratno prečno preverjanje [22, 17, 12, 11])**.** *Je oblika prečnega preverjanja, kjer podatke naključno razdelimo na k podmnožic enakih velikosti. Od teh k podmnožic izberemo eno, s katero model testiramo, drugih $k - 1$ podmnožic pa uporabimo za učenje modela. Prečno preverjanje ponovimo k-krat tako, da je vsaka podmnožica enkrat uporabljena za testiranje modela.*

Opomba: *V diplomski nalogi uporabljamo metodo 'izloči enega', ki je skrajni primer k-kratnega prečnega preverjanja, kjer je k = število primerov.*

Gradnja napovednih modelov je zelo pomembna za doseganje ciljev diplomske naloge. Na učnih podatkih smo se naučili klasificiranja (rangiranja) vzorcev in jih nato uporabili za napovedovanje fenotipov (rangov) vzorcev v testni množici. Za vsako skupino SNV-jev, ki je določena bodisi na podlagi tabele GO ali KEGG bodisi medgenskih SNV-jev, izvedemo naslednji postopek:

1. Zgradimo napovedni model tako, da za vsak gen iz izbrane skupine najdemo pripadajoče SNV-je v podatkih. Te SNV-je nato povprečimo:

$$\overline{SNV_{gen_i}} = \frac{1}{k}\sum_{j=1}^{k} SNV_{gen_i},$$

   kjer je k število SNV-jev, ki pripadajo genu gen_i.

2. Dobljeno tabelo transponiramo (SNV-ji postanejo stolpci (atributi), vzorci pa vrstice (primeri)).

3. Na transponirani podatkih opravimo 26-kratno prečno preverjanje oz. metodo 'izloči enega' (definicija 3.7) [22, 17, 12, 11].

4. Napovedno uspešnost izbrane skupine genov ocenimo z izračunom napak MAE (definicija 3.9) [20, 9] in MSE (definicija 3.8) [23, 9].

Definicija 3.8 (Povprečna kvadratna napaka ali MSE [23, 9])**.** *MSE je definirana kot:*

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(X_i - Y_i)^2,$$

kjer je X vektor z N napovedmi in Y vektor z N dejanskimi vrednosti.

Definicija 3.9 (Povprečna absolutna napaka ali MAE [20, 9])**.** *MAE je definirana kot:*

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|X_i - Y_i|,$$

kjer je X vektor z N napovedmi in Y vektor z N dejanskimi vrednosti.

Napovedne modele smo gradili tudi na podatkih, kjer smo združili posamezne skupine genov. Združevanje vseh možnih kombinacij skupin bi bilo časovno in računsko prezahtevno. Zato smo v i-tem koraku medsebojno združevali le k predhodno združenih skupin, ki so se v koraku $(i-1)$ izkazale za najbolj informativne. Trenutno najboljše kombinacije skupin smo poskusili izboljšati z dodajanjem novih skupin. Postopek je naslednji:

1. Izračunamo vse možne pare skupin genov. Če je število skupin genov n, je parov $\frac{1}{2}n(n-1)$.

2. Na podlagi SNV-jev v vsakem paru skupin zgradimo napovedni model in z napakama MAE in MSE ovrednotimo, kako dobro klasificiramo (rangiramo) vzorce.

3. Empirično določimo parameter k in vzamemo le k parov skupin genov, ki so najboljše klasificirali (rangirali) vzorce.

4. Vsak par $pair_i$; $i = 1, ..., k$ nato združimo z vsako skupino genov, ki še ni v paru. Za vsako tako združitev zgradimo napovedni model ter ga ocenimo z napakama MAE in MSE. Ker je izbranih parov k in vsakega združimo z vsemi možnimi skupinami genov, ki še niso v paru, je takih združitev $k(n-2)$.

5. Izmed vseh združitev shranimo le uspešne. To so tiste, za katere velja, da je bila napaka MAE pred dodajanjem nove skupine genov (v koraku $i-1$) večja, kot je po dodajanju skupine (v koraku i).

6. Ponovimo drugi korak, še za k najboljših združitev dolžine 3, 4, 5 itd. Število združitev je v vsakem koraku $k * (n - (\text{dolžina združitve}))$-krat.

7. Algoritem ustavimo, ko se kljub dodajanju novih skupin vrednosti napak MAE in MSE ne spreminjajo več.

Poglavje 4

Rezultati

Predstavili smo podatke in opisali metode za doseganje ciljev. V naslednjih poglavjih poročamo o uspešnosti zgrajenih napovednih modelov. Opišemo tudi uspešnost napovedovanja fenotipa na podlagi podatkov o genih, posameznih skupinah genov ter združenih skupinah genov. Za najpomembnejše gene, SNV-je in skupine genov določimo, iz katerega od staršev morajo biti podedovani, da je fenotip najbolj ugoden (odpornost na kemikalijo največja). Pokažemo tudi, koliko medgenski SNV-ji prispevajo k boljšemu napovedovanju fenotipov. Preverimo, kako dobro je možno rangirati vzorce, če fenotip diskretiziramo na nekaj vrednosti. Pokažemo, kako izbirati vzorce in koliko jih potrebujemo za izgradnjo uspešnega napovednega modela. Odgovorimo tudi na vprašanje, ali se predznak koeficientov SNV-jev in genov spreminja z napovednim modelom. Na koncu rangiramo posameznike tako, da za gradnjo napovednega modela uporabimo le vzorce IP, SP in $F1\_pool$.

4.1 Iskanje informativnih SNV-jev

V tem poglavju opišemo, kakšna je uspešnost napovedovanja fenotipa na podlagi podatkov o genih, posameznih skupinah genov ter združenih skupinah genov. Pokažemo tudi, ali večjo točnost dosežemo, če genom dodamo medgenske SNV-je.

4.1.1 Informativnost posameznih genov

Gen	#	MSE	MAE
g02911	1	25.03	4.03
g01596	3	26.73	4.11
g05774	2	35.07	4.23
g00841	1	29.92	4.3
g00930	4	38.15	4.53
g04772	3	38.88	4.96
g05782	2	46.88	4.96
g00088	2	38.15	5.0
g00627	5	42.84	5.07
g04513	1	38.96	5.19

(a) Geni dobljenih z logistično regresijo.

Gen	#	MSE	MAE
g04307	4	37.0	4.92
g04303	1	36.88	4.96
g04296	8	35.23	5.0
g04304	11	37.38	5.15
g04289	2	42.15	5.15
g03710	5	43.15	5.15
g03981	1	40.5	5.19
g04309	4	38.88	5.26
g04290	5	40.11	5.26
g03977	1	45.26	5.26

(b) Geni dobljenih z linearno regresijo.

Tabela 4.1: Deset najbolj informativnih genov.

Najprej smo preverili napovedno vrednost posameznih genov. Napovedni model smo v tem primeru zgradili na podlagi vseh SNV-jev (ki jih nismo povprečili), ki pripadajo določenemu genu. Če bi SNV-je gena povprečili, bi napovedni model uporabljal le en atribut (povprečen SNV).

Ker je genov v končnih podatkih zelo veliko, smo se odločili prikazati le deset genov, s katerimi najbolje napovemo fenotipe vzorcev. Tabela 4.1a prikazuje deset najboljših genov, dobljenih z uporabo logistične regresije. Tabela 4.1b prikazuje deset najboljših genov, dobljenih z uporabo linearne regresije. Stolpec # v tabelah pove, s koliko SNV-ji smo gradili napovedni model. Če primerjamo rezultate na obeh grafih, opazimo dve stvari:

1. napovedovanje z logistično regresijo je malenkost bolj točno,

2. najbolj informativni geni, določeni z logistično regresijo, so popolnoma drugačni od tistih, določenih z linearno regresijo. Izkaže se, da je pet

genov v preseku 100-ih najboljših genov, dobljenih z linearno regresijo in z logistično regresijo.

Vrednosti v tabelah 4.1a in 4.1b kažejo, da logistična regresija zgradi boljši napovedni model in da ne moremo pričakovati velikega preseka med množicami genov, ki ustrezajo posamezni metodi.

4.1.2 Informativnost skupin genov

Ker je genov preveč, bi bilo njihovo združevanje časovno prezahtevno. Zato smo gene združevali v skupine tako, kot smo to opisali v podpoglavju 3.1.2. Ker imamo podani dve tabeli s predznanjem (GO in KEGG), smo gene razvrstili v dve vrsti skupin.

Tabele 4.2a, 4.2b, 4.3a in 4.3b prikazujejo rezultate napovedi, če uporabimo deset najboljših skupin genov glede na tabelo predznanja (GO ali KEGG) in vrsto napovednega modela (linearno ali logistična regresija). Ker so skupine genov dobljene iz različnega predznanja (GO ali KEGG) drugačne, jih označimo z različnima imenoma. Skupine, ki so sestavljene iz genov v tabeli GO, se začnejo z oznako **c**, medtem ko se skupine, sestavljene iz genov tabele KEGG, začnejo z oznako **ck**. Stolpec # v tabelah nam pove, koliko genov je v posamezni skupini.
Ker je v vsaki skupini genov več atributov za gradnjo napovednih modelov, lahko z dobro analizo grafov potegnemo že bolj gotove zaključke.

Z linearno regresijo boljše napovedujemo fenotipe vzorcev. To je najbolj razvidno, če primerjamo tabeli 4.3a in 4.2a, pa tudi če pogledamo padec vrednosti v tabelah 4.3b in 4.2b. To sicer nasprotuje našim predpostavkam o boljšem napovedovanju fenotipov na podlagi posameznih genov.

Najbolj informativne skupine, dobljene z linearno regresijo, so precej drugačne od tistih, ki jih dobimo z logistično regresijo. Tabeli 4.3a in 4.2a nimata nobene skupne skupine genov. Tabeli 4.3b in 4.2b imata le tri skupne skupine. Predpostavka iz prejšnjega poglavja preverjeno drži.

Skup.	#	MSE	MAE
c183	6	71.34	5.8
c1	27	47.76	6.0
c185	13	58.34	6.03
c29	32	66.03	6.19
c59	12	63.38	6.23
c55	16	59.88	6.26
c136	9	71.57	6.26
c238	4	61.88	6.5
c112	16	62.57	6.5
c119	12	66.03	6.8

(a) Skupine dobljene z logistično regresijo, GO.

Skup.	#	MSE	MAE
ck5	29	55.26	5.65
ck23	4	51.53	5.84
ck79	1	74.53	6.38
ck28	42	73.46	6.69
ck37	27	75.34	6.96
ck46	5	79.73	6.96
ck62	11	75.53	7.0
ck68	10	78.15	7.07
ck27	39	75.76	7.15
ck35	9	79.5	7.34

(b) Skupine dobljene z logistično regresijo, KEGG.

Tabela 4.2: Deset najboljših skupin genov.

Če za razvrščanje genov v skupine uporabimo predznanje iz tabele GO, potem bolje napovedujemo fenotipe vzorcev. Izjema je le prva vrednost tabele 4.2b. Ob poskušanju združevanja skupin genov iz tabele KEGG v pare to trditev lahko potrdimo. Zaradi tega smo nadaljnje združevanje teh skupin genov opustili.

Primerjava vrednosti napak MAE skupin genov z vrednostmi napak MAE posameznih genov podpira nepričakovane zaključke: s posameznimi geni bolje napovedujemo fenotipe vzorcev kot s skupinami genov. Vseeno pa pričakujemo, da bomo z združevanjem skupin dobili boljše rezultate.

4.1.3 Najbolj informativne združitve skupin genov

Idejo o združevanju skupin genov smo opisali v podpoglavju 3.1.2. Tu poročamo o empiričnih rezultatih. Navajamo nabore SNV-jev, s katerimi smo najboljše napovedali fenotipe vzorcev in tako dosegli enega izmed zastavljenih ciljev.

Skup.	#	MSE	MAE
c28	33	32.26	4.88
c237	25	46.03	5.65
c188	30	57.26	5.65
c174	3	48.26	5.88
c135	4	52.11	5.88
c83	21	55.0	6.0
c75	4	59.42	6.19
c101	3	58.69	6.23
c53	12	69.69	6.23
c5	14	58.76	6.3

(a) Skupine dobljene z linearno regresijo, GO.

Skup.	#	MSE	MAE
ck63	96	58.19	6.34
ck78	1	54.5	6.42
ck79	1	57.57	6.42
ck1	63	59.65	6.5
ck6	11	59.76	6.53
ck24	42	61.34	6.65
ck71	9	61.34	6.65
ck47	3	70.57	6.73
ck46	5	67.84	6.92
ck68	10	76.5	7.03

(b) Skupine dobljene z linearno regresijo, KEGG.

Tabela 4.3: Deset najboljših skupin genov.

V tabelah 4.4 in 4.5 so prikazane najboljše združitev skupin genov glede na to, kateri napovedni model smo uporabili. Poleg tega vidimo tudi število genov v združitvah (stolpec #).

Analiza teh grafov in grafov iz prejšnjih dveh podpoglavij podpira končne zaključke o zmožnosti napovedovanja fenotipov vzorcev z uporabo predznanja.

1. Najboljše združitve skupin genov, dobljene z logistično regresijo, vsebujejo v povprečju približno trikrat več genov od najboljših združitev, dobljenih z linearno regresijo.

2. V tabelii 4.5 se največkrat pojavi skupina **c192** (4-krat), ki je šele na 49. mestu, ko fenotipe vzorcev napovedujemo le s skupinami genov. Najbolj pogoste skupine genov v drugi tabeli 4.4 pa so skupine **c55**, **c151** in **c184** (pojavijo v vseh desetih najboljših združitvah skupin genov). Ko napovedujemo samo s posameznimi skupinami, je skupina c55 na 6. mestu, skupina c151 na 117. mestu, skupina c184 pa šele na

Združitev skupin	#	**MSE**	**MAE**
c55, c73, c114, c151, c184, c202, c219	93	8.61	2.46
c13, c55, c73, c151, c184, c202, c213	92	10.07	2.61
c31, c55, c73, c114, c151, c184, c202	104	10.65	2.65
c31, c55, c73, c151, c184, c202	85	10.15	2.69
c55, c73, c151, c184, c200, c202, c223	79	10.46	2.69
c13, c55, c73, c151, c164, c184, c202	86	10.65	2.73
c55, c73, c151, c163, c174, c184, c202	81	10.61	2.76
c31, c151, c157, c169, c184, c209	41	11.53	2.76
c31, c55, c101, c151, c184, c202, c228	78	11.69	2.76
c31, c55, c73, c151, c184, c201	101	12.46	2.76

Tabela 4.4: Deset najbolj informativnih združitev skupin genov, dobljenih z uporabo logistične regresije.

138. mestu.

Zaključimo lahko, da nekatere skupine genov zelo dobro sovpadajo s točno določenimi drugimi skupinami genov.

3. Napovedovanje fenotipov vzorcev z linearno regresijo je precej bolj natančno. Z logistično regresijo smo napovedovali vzorce fenotipov zato, ker smo najboljše dobljene združitve skupin genov potrebovali pri rangiranju vzorcev.

4.1.4 Souporaba skupin genov in medgenskih SNV-jev

V razdelku 3.1.3 smo opisali razloge za uporabo medgenskih SNV-jev. Pričakovali smo, da bodo nosili dodatno informacijo o fenotipu in tako še izboljšali napovedno točnost.

Najprej smo poskusili tako, da smo najboljšim dobljenim združitvam skupin genov dodali vse možne skupine medgenskih SNV-jev. Ker se rezultati tako niso izboljšali, smo se odločili poskusiti še s tem postopkom:

Združitev skupin	#	**MSE**	**MAE**
c155, c164, c171, c192, c222	24	1.08	0.62
c132, c171, c192, c218	23	0.92	0.69
c135, c193	24	1.46	0.69
c4, c134, c191, c218, c228	24	1.04	0.73
c5, c121, c132, c132, c164, c200	24	1.35	0.73
c53, c155, c192, c209	25	1.23	0.77
c150, c190, c222	25	1.58	0.81
c4, c134, c191, c228	23	1.31	0.85
c43, c53, c190, c192, c200	30	1.65	0.89
c53, c155, c192, c238	24	2.42	1.04

Tabela 4.5: Deset najbolj informativnih združitev skupin genov, dobljenih z uporabo linearne regresije.

1. Pogledamo, s katerimi skupinami medgenskih SNV-jev cmg_i najbolje napovedujemo vzorce fenotipov. Vzamemo le tiste, za katere velja $MAE(cmg_i) < 7.0$. Le 8 skupin medgenskih SNV-jev zadošča temu kriteriju.

2. Združimo vsako skupino genov z vsako izmed skupin medgenskih SNV-jev. Ker je število skupin genov iz datoteke GO enako 240 in smo vzeli le 8 najboljši skupin medgenskih SNV-jev, je možnih združitev $240 * 8 = 1920$. Za vsako izmed teh združitev smo zgradili napovedni model. Nato smo opravili 26-kratno prečno preverjanje in vzporedno napovedali fenotipe vzorcev. Njihovo točnost smo ocenili z napakama MAE in MSE.

3. Med temi združitvami za nadaljnje združevanje smo uporabili le najboljših k (k je določen empirično) združitev, ker bi sicer bilo združevanje časovno prezahtevno.

4. Od tu naprej je postopek enak tistemu, ki smo ga uporabili za združevanje skupin genov.

Združitev skupin	#	MSE	MAE
c15, c54, c151, c183, c200, mg70	69	9.69	2.61
c13, c54, c151, c169, mg70	65	10.57	2.65
c15, c54, c151, c183, c228, mg70	69	11.03	2.73
c15, c54, c56, c151, c202, mg70	89	12.53	2.76
c24, c72, c120, c145, c168, mg14	57	10.73	2.8
c24, c183, c191, c208, c225, mg14	58	11.34	2.8
c15, c54, c89, c151, c183, mg70	82	11.88	2.8
c15, c54, c151, c169, c228, mg70	66	12.42	2.8
c24, c72, c183, c191, c208, mg14	56	12.65	2.8
c8, c15, c54, c151, c206, mg70	78	11.53	2.84

Tabela 4.6: Deset najboljših združitev skupin genov s souporabo skupin medgenskih SNV-jev, dobljenih z logistično regresijo.

S postopkom pridemo do združitev skupin, kjer vsaka vsebuje tudi eno skupino medgenskih SNV-jev. Najboljše izmed njih prikažemo v tabelah 4.6 in 4.7. Stolpec # v tabelah pove, koliko genov in povprečnih medgenski SNV-jev je v združitvi skupin. Z dodajanjem medgenskih SNV-jev smo hoteli izboljšati natančnost napovedovanja fenotipov vzorcev.

Primerjava tabel 4.6 in 4.7 s tabelama 4.4 in 4.5 pokaže, da najboljši rezultat dobimo s souporabo skupin medgenskih SNV-jev. Vendar pa ob primerjavi povprečnih napak na grafih ugotovimo, da smo boljše rezultate dosegli brez uporabe skupin medgenskih SNV-jev.

Pri rangiranju vzorcev se izkaže, da z združitvami skupin s souporabo medgenskih SNV-jev dobimo veliko slabše rezultate. Zaključimo lahko, da medgenski SNV-jev ne nosijo veliko dodatne informacije o fenotipu.

4.2 Odkriti geni in SNV-ji

V prejšnjem poglavju smo navajali združitve skupin genov, ki najbolje napovedujejo fenotipe vzorcev. Vsaka izmed njih vsebuje več genov, ki pa so

Združitev skupin	#	MSE	MAE
c118, c148, c218, mg94	24	0.65	0.57
c118, c148, c218, c228, mg94	25	0.76	0.61
c0, c75, mg94	24	2.46	0.92
c62, c118, c148, c200, mg94	28	2.61	1.0
c125, c132, mg94	24	3.0	1.07
c118, c132, c148, c200, mg94	26	4.5	1.19
c118, c148, c164, c218, c228,mg94	26	3.42	1.26
c62, c118, c148, mg94	27	3.65	1.26
c75, c118, c134, c174, c190, mg94	31	2.69	1.3
c118, c148, c209, mg94	25	4.62	1.39

Tabela 4.7: Deset najboljših združitev skupin s souporabo skupin medgenskih SNV-jev, dobljenih z linearno regresijo.

sestavljeni iz večih SNV-jev. V tem poglavju pokažemo, kateri od teh genov (SNV-jev) so bolj informativni za napovedovanje fenotipov vzorcev.

4.2.1 Logistična regresija

Zanimalo nas je, na kakšen način določiti povezanost gena (SNV-ja) s fenotipom. Logistična regresija deluje tako, da vsakemu primeru s_i v napovednem modelu priredi seznam koeficientov

$$[k_{s_i,a_1}, k_{s_i,a_2}, ..., k_{s_i,a_m}]$$

kjer je m število atributov a. V naših podatkih so primeri vzorci in atributi geni (SNV-ji). Pogledali smo vrednosti koeficientov atributov pri skrajnih vrednostih fenotipa. Minimalno vrednost fenotipa ima vzorec F7JP_01 ($F_{F7JP\_01} = 1$), maksimalno pa ima vzorec IP ($F_{IP} = 26$). Najbolj informativni atributi a_j so tisti, ki imajo koeficienta k_{IP,a_j} in k_{F7JP_01,a_j} čim bolj različna. Izračunati moramo torej absolutno razliko $aDiff_{a_j}$ med koeficientoma k_{IP,a_j} in k_{F7JP_01,a_j} za vsak atribut a_j:

$$aDiff_{a_j} = |k_{IP,a_j} - k_{F7JP\_01,a_j}|.$$

Gen	$coef_{F7JP\_01}$	$coef_{IP}$	**skupina**	**Absolutna razlika**
g00924	0.0870	-0.1403	clu219	0.2273
g02968	-0.0156	-0.2091	clu73	0.1935
g02434	-0.0567	-0.2354	clu151	0.1786
g01203	-0.0419	-0.2143	clu114	0.1724
g04704	0.1017	-0.070	clu73	0.1719
g00136	0.1419	-0.0234	clu184	0.1653
g03505	0.2207	0.0646	clu73	0.1562
g04338	0.2386	0.0897	clu73	0.1489
g02988	-0.0628	-0.2075	clu73	0.1446
g01028	0.0938	-0.0478	clu55	0.1416

Tabela 4.8: Deset najbolj informativnih genov v najboljši združitvi skupin genov, dobljeni z logistično regresijo.

Atribute smo nato uredili po padajoči absolutni razliki in jih tako razvrstili od najbolj pomembnega do najmanj pomembnega. V tabeli 4.8 je predstavljenih deset najbolj pomembnih genov v najboljši združitvi skupin genov, dobljeni z logistično regresijo. Ta združitev skupin genov je sestavljena iz: $c73$, $c151$, $c184$, $c202$, $c55$, $c219$ in $c114$. Rezultati v tabele pokažejo, da je najbolj pomembna skupina v združitvi $c73$, saj tej skupini pripada kar pet od desetih najbolj pomembnih genov.

Poglejmo še deset najbolj pomembnih SNV-jev v isti združitvi skupin genov (tabela 4.9. Za pričakovati je, da bodo vsi ali pa vsaj večina SNV-jev pripadali enemu izmed genov v tabeli 4.8. Vidimo, da razen SNV-ja snv47810, ki pripada genu g04097, vsi drugi SNV-ji pripadajo enemu izmed desetih najbolj pomembnih genov. Opazimo še, da prvi trije SNV-ji pripadajo najbolj pomembnemu genu g00924. To samo še potrdi pomembnost gena v tej združitvi skupin genov.

SNV	**Gen**	$coef_{F7JP\_01}$	$coef_{IP}$	**Absolutna razlika**
snv11464	g00924	0.0553	-0.0231	0.0784
snv11466	g00924	0.0475	-0.0280	0.0755
snv11469	g00924	0.0255	-0.0420	0.0675
snv50385	g04338	0.0781	0.0204	0.0578
snv13722	g01203	0.0034	-0.0532	0.0565
snv50380	g04338	0.0599	0.0035	0.0564
snv42206	g03505	0.0538	-0.0026	0.0564
snv13721	g01203	0.0222	-0.0322	0.0543
snv47810	g04097	0.0202	-0.0341	0.0543
snv02050	g00136	0.0547	0.0015	0.0532

Tabela 4.9: Deset najbolj informativnih SNV-jev v najboljši združitvi skupin genov, dobljeni z logistično regresijo.

4.2.2 Linearna regresija

Določanje povezanosti gena (SNV-ja) s fenotipom je pri linearni regresiji drugačna. Za razliko od logistične regresije linearna regresija priredi vsakemu atributu a_j samo en koeficient k_{a_j}, ne glede na število primerov v napovednem modelu. Če je število atributov enako m, potem dobimo enodimenzionalni seznam koeficientov:

$$[k_{a_1}, k_{a_2}, ..., k_{a_m}] .$$

Atribute lahko potem od najbolj do najmanj pomembnega razvrstimo tako, da jih sortiramo padajoče po absolutni vrednosti koeficienta.

Tabela 4.10 navaja deset najbolj informativnih genov v najbolj informativni združitvi skupin genov, dobljeni z linearno regresijo. Ta združitev skupin genov je sestavljena iz skupin: $c171$, $c192$, $c222$, $c155$ in $c164$. Najbolj pomembni skupini sta $c171$ in $c192$, saj vsi geni izmed desetih najbolj pomembnih pripadajo eni izmed teh dveh skupin.

Poglejmo še deset najbolj pomembnih SNV-jev v tej združitvi skupin

Gen	**skupina**	$coef_{gen_i}$
g04583	clu171	-13.2115
g05221	clu192	-10.8298
g04895	clu192	9.8542
g05669	clu171	8.9917
g04651	clu192	8.9545
g03948	clu171	-7.0020
g04973	clu192	-6.4913
g05705	clu192	-6.3552
g01135	clu171	-6.2073
g05039	clu171	-6.1384

Tabela 4.10: Deset najbolj informativnih genov v najboljši združitvi skupin genov, dobljeni z linearno regresijo.

genov. Tudi tukaj je za pričakovati, da bodo vsi ali pa vsaj večina od desetih SNV-jev v tabeli 4.11 pripadali enemu od desetih najbolj pomembnih genov v tabeli 4.10. Vendar se to ne zgodi. Kar šest od desetih najpomembnejših SNV-jev ne pripada nobenemu genu iz tabele 4.10.

Zanimivo je tudi, da imata koeficienta SNV-jev snv57975 in snv57974, ki pripadata istemu genu, nasprotni predznak. To pomeni, da morata biti med seboj zelo različna. Če pogledamo v množico končnih podatkov, ugotovimo, da se res precej razlikujeta, in sicer kar pri dvanajstih vzorcih.

4.2.3 Podedovani SNV-ji in fenotip

Eden izmed ciljev diplomske naloge je najti način, s katerim bomo SNV-je določili, od katerega starša morajo biti podedovani, da je fenotip vzorca dober. To lahko naredimo z uporabo koeficientov v tabelah 4.8, 4.9, 4.10 in 4.11. Ker smo uporabili dva različna napovedna modela, bomo za vsakega definirali drugačen način določanja:

Pri genih (SNV-jih), dobljenih z uporabo *logistične regresije* velja, da bo

SNV	**Gen**	$coef_{SNV_i}$
SNV04546	g00341	-1.7861
SNV05481	g00419	1.7405
SNV09416	g00714	-1.5790
SNV57975	g04973	-1.3787
SNV54035	g04651	1.1909
SNV54030	g04651	1.1909
SNV57974	g04973	1.018
SNV46144	g03842	1.018
SNV04326	g00323	-1.0050
SNV60952	g05201	0.9914

Tabela 4.11: Deset najbolj informativnih SNV-jev v najboljši združitvi skupin genov, dobljeni z linearno regresijo.

fenotip vzorca dober, če bo gen_i (SNV_i) podedovan od starša:

$$\begin{cases} SP, & coef_{F7JP\_01,SNV_i} > 0 \ \wedge \ coef_{IP,SNV_i} < 0 \\ IP, & coef_{F7JP\_01,SNV_i} < 0 \ \wedge \ coef_{IP,SNV_i} > 0 \\ undefined, & \text{sicer.} \end{cases}$$

Od katerega starša morajo biti podedovani geni (SNV-ji), dobljeni z uporabo *linearne regresije*, da je fenotip vzorca dober, pa določimo tako:

$$\begin{cases} SP, & coef_{SNV_i} < 0 \ (coef_{gen_i} < 0) \\ IP, & \text{sicer.} \end{cases}$$

Pričakovati je, da bo večina genov (SNV-jev) podedovanih od večvrednega starša (SP), saj je njegov fenotip precej nižji od fenotipa manjvrednega starša (IP). Zaradi tega mora biti tudi vsota koeficientov, dobljenih z linearno regresijo, negativna. V tabeli 4.12 se lahko prepričamo, da sta naši trditvi pravilni.

Model	geni ali SNV-ji	# SP	# IP	# 'undet'	vsota
log. reg.	geni	20	2	71	/
log. reg.	SNV-ji	121	34	330	/
lin. reg	geni	12	12	/	-8.001
lin. reg	SNV-ji	66	43	/	-8.000

Tabela 4.12: Število SNV-jev, ki morajo biti podobni SP-ju oz. IP-ju, da bo fenotip nizek.

4.3 Funkcijski pripisi odkritih genov

V prejšnjih poglavjih smo dobili združitve skupin genov, ki najbolje napovedujejo fenotipe vzorcev. Do osnovnih skupin genov pa smo prišli tako, da smo v isto skupino razvrstili gene, ki sodelujejo pri istih procesih. V tem poglavju nas zanima, s katerimi pripisi GO lahko opišemo gene v najboljših združitvah grup in kolikšna je verjetnost, da bi te gene dobili naključno. To bomo merili s p-vrednostjo (definicija 4.1 [25, 7]) in deležem napačno pozitivnih zadetkov ali **FDR** (definicija 4.2 [6, 29, 1]). Zaradi dobrih implementacij smo se odločili, da bomo za realizacijo opisanih stvari uporabili razred *Orange.bio* iz programskega orodja *Orange* [3].

Definicija 4.1 (P-vrednost [25, 7])**.** *P-vrednost je definirana kot verjetnost (pod neko hipotezo **H**), da dobimo rezultat, ki je enak ali bolj ekstremen kot dobljeni rezultat. Z njo ovrednotimo moč dokazov proti ničelni hipotezi in za hipotezo **H**. Če je p-vrednost zelo majhna (po navadi $p < 0.01$), to pomeni, da smo dobili močne dokaze proti ničelni hipotezi.*

Definicija 4.2 (Delež napačno pozitivnih zadetkov ('False Discovery Rate' ali **FDR**) [6, 29, 1])**.** *Naj bo $P^m = (P_1, ..., P_m)$ seznam p-vrednosti za m testov in $P_{(0)} \equiv 0$. Sortiramo p-vrednosti tako, da velja:*

$$P_{(0)} = 0 < P_{(1)} < ... < P(_{(m)}.$$

Določimo še indikatorje hipotez $H^m = (H_1, .., H_m)$, kjer je $H_i = 0$, če je

Pog.	GO_id:	Izraz GO	p	FDR	rcm	gcm
P	0010556:	reg. of macromolecule biosynthetic p.	1.14e-26	3.6e-24	711	63
P	0044249:	cellular biosynthetic p.	1.27e-23	1.34e-21	1562	81
C	0030134:	ER to Golgi transport vesicle	4.57e-08	3.05e-06	27	10
F	0002161:	Aminoacyl-tRNA editing activity	7.16e-07	9.6e-05	6	6
C	0044427:	chromosomal part	4.6e-08	3.05e-06	314	26
F	0042162:	telomeric DNA binding	1.02e-06	9.6e-05	20	8
C	0033588:	Elongator holoenzyme complex	2.29e-06	2.7e-05	6	5
C	0005634:	nucleus	2.29e-06	2.7e-05	1761	61

Tabela 4.13: Funkcijski pripisi genov iz najboljše združitve skupin dobljene z logistično regresijo.

*pravilna i-ta ničelna hipoteza, in $H_i = 1$, če je pravilna i-ta hipoteza H. Definiramo enačbo za izračun deleža napačno pozitivnih odkritij ('False Discovery Proportion' ali **FDP**) za prag t, in sicer tako:*

$$FDP(t) = \frac{\sum_i 1\{P_i \leq t\}(1 - H_i)}{\sum_i (1\{P_i \leq t\} + 1\{P_i < t\})}.$$

*Delež napačno pozitivnih odkritij (**FDR**) za več testnih pragov T definiramo kot matematično upanje za $FDP(T)$.*

V drevesu GO-izrazov obstajajo trije pogledi pripisov GO: biološki procesi, predeli celice in molekularna funkcija. Za najboljše združitve skupin genov smo poiskali pripise GO za vsak pogled. Ker je teh veliko, se odločimo, da bomo prikazali le najboljšega (z najmanjšo p-vrednostjo in FDR-jem) za vsak pogled (tabeli 4.13 in 4.14). Povejmo še, da je število možnih genov (*refe-*

renca) 4972 in da sta preiskovana nabora genov dolga 24 genov za linearno in 93 genov za logistično regresijo. Pomen stolpcev v tabelah je naslednji:

1. **Pogled** ('Aspect') pove, katere vrste pripisov GO preiskujemo. Vrednosti so lahko 'P' (biološki proces), 'C' (predel celice) ali 'F' (molekularna funkcija).

2. **P**-vrednost nam pove, kolikšna je verjetnost po hipergeometricni distribuciji (naključno žrebanje brez vračanja), da z naključnim izbiranjem genov dobimo preiskovan nabor genov. To pomeni, da izberemo **gcm** genov od preiskovanega nabora genov (24 ali 93), ki pripadajo dobljenemu **izrazu GO**, če je v *referenci* takšnih genov **rcm**.

3. **FDR** je po definiciji 4.2 popravljena **p**-vrednost.

4. **GO_id** je unikatna identifikacijska številka **izraza GO**.

5. **Izraz GO** nam pove ime biološkega procesa, predela celice ali molekularne funkcije.

6. Število **rcm** nam pove, koliko genov iz *reference* sodeluje pri dobljenem izrazu GO.

7. Število **gcm** nam pove, koliko genov iz preiskovange nabora genov sodeluje pri dobljenem izrazu GO.

4.4 Referenčne vrednosti

Za vrednotenje dobljenih rezultatov potrebujemo referenčne vrednosti. Te vrednosti dobimo na različne načine in tako preverimo smiselnost vsakega koraka postopka. Rezultate, pridobljene s predlaganimi postopki, primerjamo z referenčnimi rezultati in tako lahko z večjo gotovostjo ocenimo uspešnost postopka.

Pog.	GO_id:	Izraz GO	p	FDR	rcm	gcm
P	0009891:	positive regulation of biosynthetic p.	1.42e-22	3.62e-20	278	22
C	0033698:	Rpd3L complex	9.82e-14	1.1e-11	11	8
C	0000118:	histone deacetylase complex	1.24e-12	3.46e-11	31	9
F	0045182:	translation regulator activity	2.89e-11	2.68e-09	12	7
C	0044451:	nucleoplasm part	2.44e-08	5.47e-07	190	11
C	0044422:	organelle part	6e-07	6.72e-06	2249	24
F	0033558:	protein deacetylase activity	2.16e-05	0.000668	13	4
F	0043022:	ribosome binding	3.73e-05	0.000694	17	4

Tabela 4.14: Funkcijski pripisi genov iz najboljše združitve skupin, dobljene z linearno regresijo.

Postopek	Nap. model	# atributov	MAE	MSE
predznanje	log. reg.	93	2.46	8.62
vsi SNV-ji	log. reg.	29905	8.16	84.12
vsi geni	log. reg.	4972	7.92	94.0
predznanje	lin. reg.	24	0.62	1.08
vsi SNV-ji	lin. reg.	29905	7.0	67.77
vsi geni	lin. reg.	4972	7.04	74.12

Tabela 4.15: Primerjava točnosti napovedovanja z vsemi geni (SNV-ji) in s postopkom izbiranja genov na podlagi predznanja.

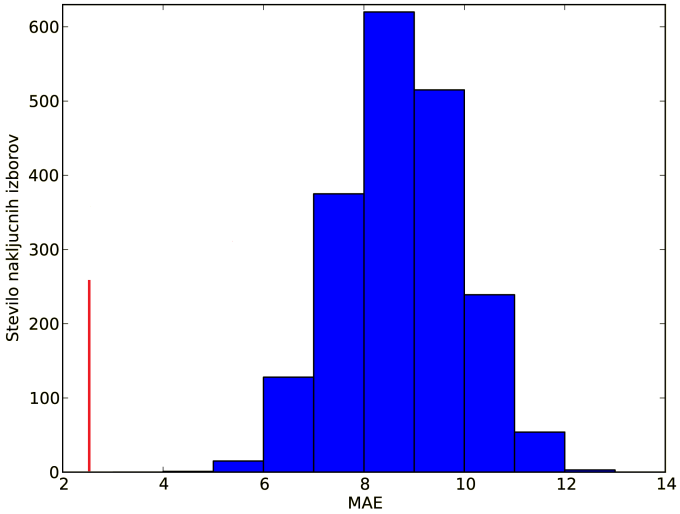
V naslednjih poglavjih pokažemo, kako smo pridobili več različnih referenčnih vrednosti. Vsako izmed njih smo primerjali z rezultati, ki smo jih dobili po postopku, tj. na osnovi predznanja.

4.4.1 Napovedna točnost celotne podatkovne baze

Zanima nas, kako dobro napovemo fenotipe vzorcev, če za gradnjo učnega modela uporabimo kar vse gene (SNV-je). Rezultati v tabeli 4.15 kažejo, da dobimo z izbiranjem genov na osnovi predznanja (GO in KEGG) boljše rezultate, kot če za napovedovanje fenotipov vzorcev uporabimo vse gene (SNV-je). To je prvi dokaz, da je naš postopek pravilen.

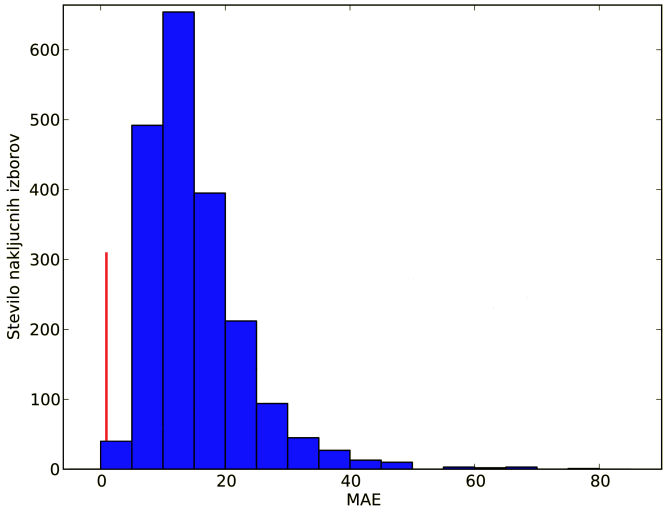
4.4.2 Naključen izbor genov

Pokazali smo, da z izbiranjem genov na osnovi predznanja dobimo boljše rezultate, kot če gene izbiramo naključno. Pri napovedovanju fenotipov vzorcev z različnim napovednim modelom smo dobili različno velike nabore genov (SNV-jev). Pri uporabi logistične regresije smo v povprečju najboljše rezultate dosegli z združitvami skupin genov, ki so bile sestavljene iz približno 100 genov. Pri uporabi linearne regresije pa so bile najboljše združitve skupin genov v povprečju sestavljene iz okoli 30 genov. Zaradi tega smo pri napovedovanju fenotipov vzorcev z logistično regresijo naključno izbirali po 100



Slika 4.1: Napaka MAE predlaganega postopka z uporabo logistične regresije je označena z rdečo, navpično črto in znaša 2.46.

genov, pri napovedovanju z linearno regresijo pa 30 genov.



Slika 4.2: Napaka MAE predlaganega postopka z uporabo linearne regresije je označena z rdečo, navpično črto in znaša 0.62.

Ker nas pri naključnih postopkih zanima predvsem, kako dobri so v povprečju, smo gene naključno izbrali večkrat (2000-krat). Na slikah 4.2 in 4.1

Postopek	**Nap. model**	**Rezultat**	**MAE**	**MSE**	**P(MAE)**
predznanje	log. reg.	najboljši	2.46	8.62	0%
naključen	log. reg.	povprečen	8.76	110.8	47.4%
naključen	log. reg.	najboljši	4.85	34.31	0.0005%
predznanje	lin. reg.	najboljši	0.62	1.08	0%
naključen	lin. reg.	povprečen	15.18	379.6	54.85%
naključen	lin. reg.	najboljši	2.81	12.73	0.0005%

Tabela 4.16: Primerjava napak MAE pri napovedovanju fenotipa vzorcev z naključnim izborom genov in z izborom genov na osnovi predznanja.

lahko vidimo porazdelitev napak MAE naključnih izborov genov. Opazimo, da lahko tudi z naključnim izbiranjem izberemo nabor genov, s katerim dobimo kar dober rezultat. Zaradi tega smo najbolj pomembne rezultate primerjali še v tabeli 4.16. V stolpcu **P(MAE)** lahko vidimo, kolikšna je verjetnost, da z naključnim postopkom dobimo manjšo ali enako napako MAE (stolpec **MAE**).

Iz podatkov v tabeli 4.16 je razvidno, da so tudi najboljši rezultati, dobljeni z naključnim izborom genov, precej slabši od rezultatov izbiranja genov na osnovi predznanja. Zato z gotovostjo zaključimo, da je naš način izbiranja genov bistveno boljši od naključnega.

4.4.3 Povprečni vektorji skupin

Tudi v tem razdelku smo preverjali smiselnost razvrščanja genov v skupine na osnovi predznanja (GO in KEGG). Dobljene skupine genov smo nato med seboj združevali in tako dobili najbolj informativen nabor genov (SNV-jev). Preverili smo, ali je bolje napovedati fenotipe vzorcev s povprečnimi predstavniki skupin. To smo naredili po naslednjem postopku:

1. Uporabimo iste skupine genov, dobljene na osnovi predznanja (GO ali KEGG) kot pri našem postopku.

2. Za vsako skupino $group_i$ inicializiramo pripadajoč ničeln vektor $\vec{group_i} = [0, 0, ..., 0, 0]$ dolžine števila vzorcev.

3. Vsak SNV *prištejemo* tistemu vektorju skupine, ki mu glede na razvrščanje genov v skupine pripada:

$$\vec{group_i} = \vec{group_i} + \vec{SNV_j};\ SNV_j \in gen_k \in group_i$$

Vzporedno štejemo, koliko SNV-jev spada v določeno skupino:

$$count_{group_i} = count_{group_i} + 1;\ SNV_j \in gen_k \in group_i$$

To počnemo, ker smo iz začetnih podatkov odstranili precej SNV-jev. Zaradi tega ne vemo točno, koliko je SNV-jev spada v katero skupino.

4. Da dobimo povprečne vektorje skupin genov $\vec{ave_{group_i}}$, jih moramo še *deliti* s številom SNV-jev, ki smo jih v tem vektorju sešteli:

$$\vec{ave_{group_i}} = \frac{1}{count_{group_i}} \vec{group_i}.$$

5. Z dobljenimi povprečnimi vektorji skupin genov zgradimo napovedni model. Opravimo 26-kratno prečno preverjanje in vzporedno z izbranim napovednim modelom napovedujemo fenotipe vzorcev.

6. Izračunamo napaki MAE in MSE ter tako ocenimo napovedno uspešnost modela.

Rezultati, dobljeni z našim in zgornjim postopkom, so prikazani v tabeli 4.17. Iz nje lahko ugotovimo, da fenotipe vzorcev po našem postopku bolje napovedujemo že z uporabo ene same skupine. Zaradi tega lahko zaključimo, da je izbiranje skupin genov na naš način zagotovo boljše.

Datoteka	Postopek	Reg. model	MAE	MSE
GO	predznanje	log. reg.	5.81	71.34
GO	povprečen	log. reg.	8.04	91.88
KEGG	predznanje	log. reg.	5.65	55.27
KEGG	povprečen	log. reg.	9.04	125.65
GO	predznanje	lin. reg.	4.89	32.27
GO	povprečen	lin. reg.	6.35	57.42
KEGG	predznanje	lin. reg.	6.35	58.19
KEGG	povprečen	lin. reg.	11.5	183.65

Tabela 4.17: Primerjava napak MAE pri napovedovanju fenotipa vzorcev z izbiranjem skupin genov po našem postopkom in z izbiranjem povprečnega predstavnika vsake skupine.

4.4.4 Kartezični produkt skupin genov

Dokazati smo želeli, da informativnost skupine ne raste, če poskušamo na umeten način povečati število elementov v njej. Število genov v skupini smo povečali tako, da smo zmnožili vse pare genov med seboj. Izračunali smo torej vse kartezične produkte skupin genov. Postopek smo ponovili za oba napovedna modela (logistično in linearno regresijo) in za obe tabeli predznanja (GO in KEGG). Postopek, po katerem smo to naredili, je naslednji:

1. Uporabimo iste skupine genov, dobljene na osnovi predznanja (GO ali KEGG) kot pri našem postopku.

2. Za vsak gen iz izbrane skupine najdemo pripadajoče SNV-je v končnih podatkih. Te SNV-je nato spovprečimo tako:

$$\overline{SNV_{gen_i}} = \frac{1}{k} \sum_{j=1}^{k} SNV_{gen_i},$$

kjer je k število SNV-jev, ki pripadajo genu gen_i.

3. Izračunamo vse možne pare dobljenih povprečnih SNV-jev $\overline{SNV_{gen_i}}$. Če je število takšnih SNV-jev v skupini n, je število možnih parov enako $\frac{n(n-1)}{2}$.

4. Za vsak par povprečenih SNV-jev izračunamo njun kartezični produkt $CP_{\overline{SNV_{gen_i}},\overline{SNV_{gen_j}}}$; $i \neq j$ in tako dobimo $\frac{n(n-1)}{2}$ različnih kartezičnih produktov:

$$CP_{\overline{SNV_{gen_i}},\overline{SNV_{gen_j}}}[k] = \overline{SNV_{gen_i}}[k] \times \overline{SNV_{gen_j}}[k].$$

5. Z izračunanimi kartezičnimi produkti povprečnih SNV-jev izgradimo napovedni model. Opravimo 26-kratno prečno preverjanje in vzporedno z izbranim napovednim modelom napovedujemo fenotipe vzorcev.

6. Izračunamo napako MAE in z njo ovrednotimo točnost napovedovanja fenotipov vzorcev.

Pričakovali smo, da se napake MAE za isto skupino ne bodo zelo razlikovale glede na uporabljeni postopek (naš ali kartezični produkt). Uspešnost postopka smo merili tako, da smo prešteli, pri koliko skupinah genov se je napaka MAE zmanjšala z uporabo kartezičnega produkta. Spomnimo, da so geni iz tabele GO razvrščeni v 240 skupin, geni iz tabele KEGG pa v 80 skupin.

Linearna regresija, GO. Pri uporabi linearne regresije na skupinah genov iz GO se pri kartezičnem produktu povprečenih SNV-jev napaka MAE zmanjša le pri 6 od 240 skupin.

Logistična regresija, GO. Število skupin, pri katerih se napaka MAE zmanjša z uporabo kartezičnega produkta, je kar 81 (dobra tretjina vseh skupin).

Logistična regresija, KEGG. Število skupin, kjer se pri teh parametrih napaka MAE zmanjša ob uporabi kartezičnega produkta, je 23 (manj kot tretjina).

Linearna regresija, KEGG. Tudi tukaj je število skupin, kjer se napaka MAE zmanjša ob uporabi kartezičnega produkta, majhno (le 5).

Po analizi teh rezultatov lahko z gotovostjo trdimo, da z uporabo kartezičnega produkta v skupinah ne izboljšamo natančnosti napovedovanja fenotipov vzorcev. Dokazali smo, da umetno povečevanje elementov v skupini v povprečju ne povečuje informativnosti.

4.5 Pomen koeficientov SNV-jev v napovednem modelu

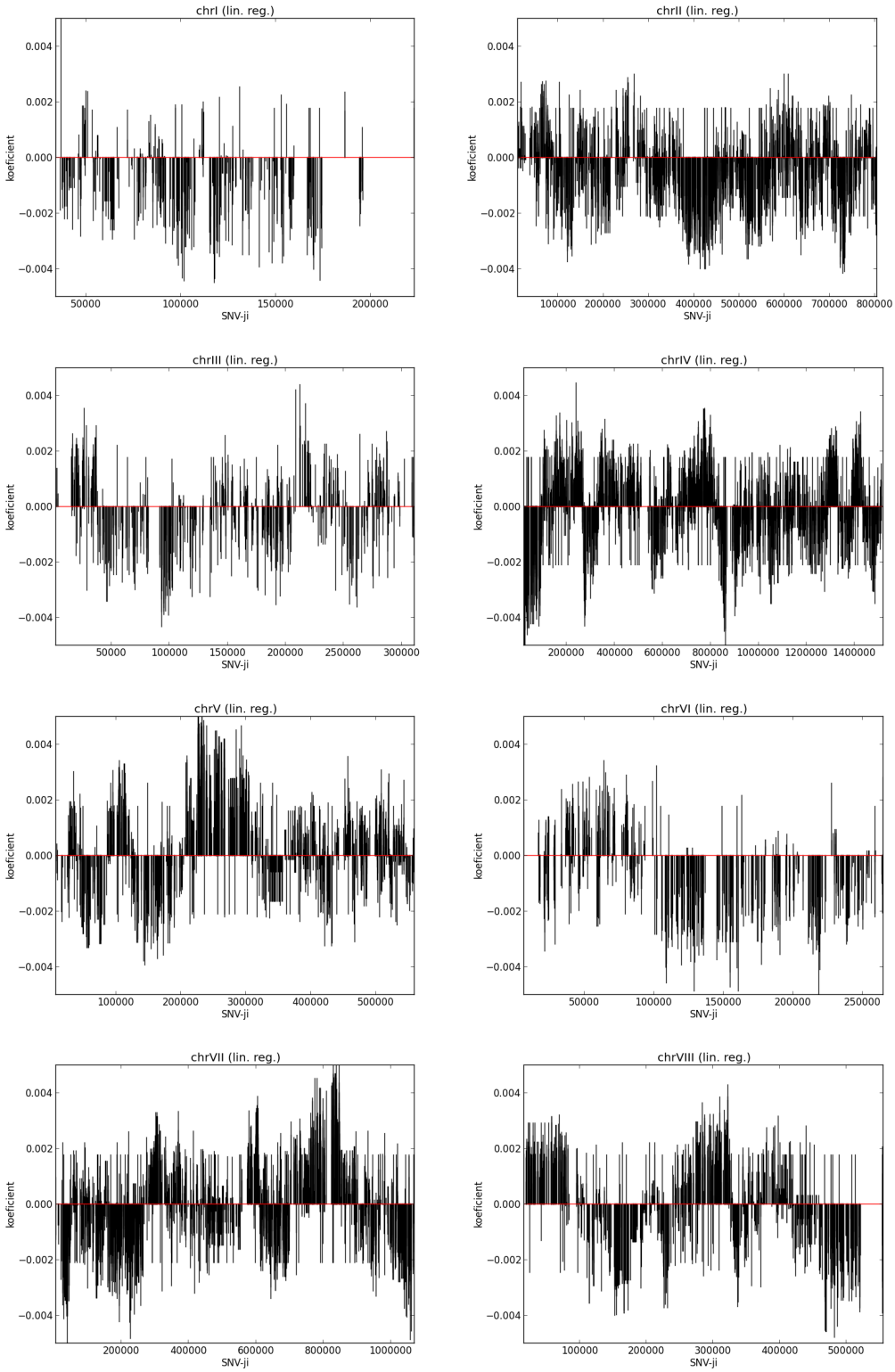
S koeficienti smo se srečali že v poglavju 4.2. V tem poglavju pa smo skušali odgovoriti na vprašanje, ali so predznaki koeficientov SNV-jev odvisni od izbora napovednega modela. Najprej smo izrisali vrednosti koeficientov SNV-jev modela, kjer smo uporabili vse SNV-je (29905 SNV-jev). Nato smo izrisali še grafe koeficientov SNV-jev modelov, ki bo bili zgrajeni iz najboljše dobljene združitve skupin genov. Ti dve skupini grafov smo nato primerjali.

Ker je kromosomov 17, smo za vsak napovedni model (linearna ali logistična regresija) in za obe vrsti napovednega modela izrisali 17 grafov.

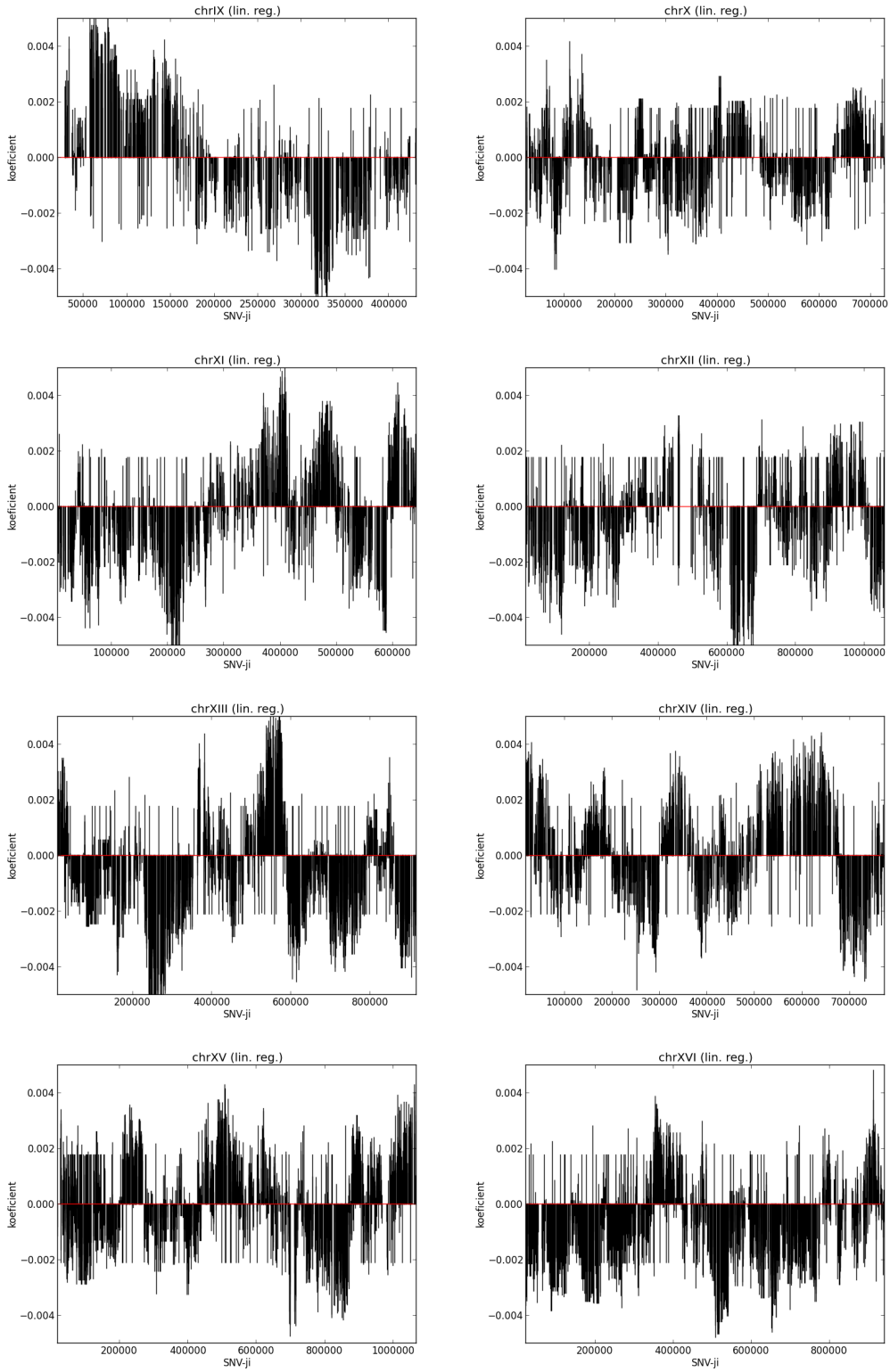
4.5.1 Linearna regresija

Poglejmo, ali se predznaki koeficientov SNV-jev, dobljenih z linearno regresijo, spreminjajo ob izbiri drugačnega napovednega modela. Primerjajmo grafe na slikah 4.3, 4.4 in 4.5 sz tistimi na slikah 4.6, 4.7 in 4.8.

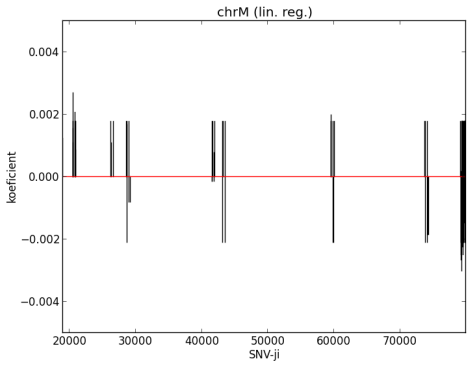
Najboljša združitev skupin genov je sestavljena le iz 24 genov oziroma 109 SNV-jev. Največ SNV-jev iz opazovane združitve skupin genov pripada drugemu kromosomu (chrII). Zato primerjamo drugi graf na sliki 4.6 z drugim grafom na sliki 4.3. Zdi se, da se koeficienti iz najboljše združitve skupin prekrivajo s koeficienti vseh SNV-jev iz končnih podatkov. Če primerjamo grafa 15. in 16. kromosoma (chrXV in chrXVI), pridemo do istega zaključka.



Slika 4.3: Vrednosti koeficientov, ko uporabimo vse SNV-je in linearno regresijo (chrI-chrVIII).



Slika 4.4: Vrednosti koeficientov, ko uporabimo vse SNV-je in linearno regresijo (chrIX-chrXVI).



Slika 4.5: Vrednosti koeficientov, ko uporabimo vse SNV-je in linearno regresijo (chrM).

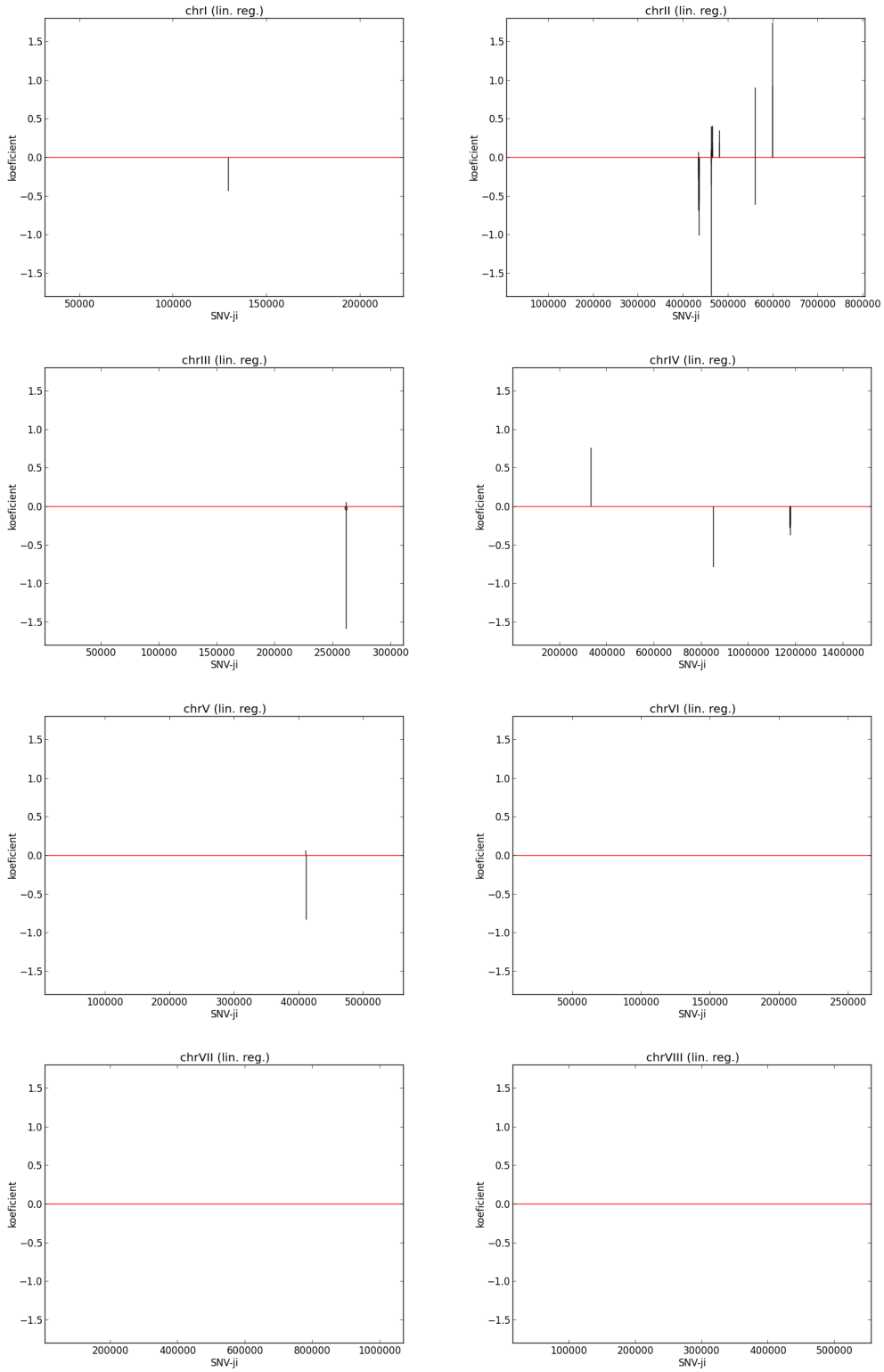
Zaradi majhnega števila SNV-jev v opazovani združitvi skupin genov gotovih zaključkov ne moremo potegniti. Na osnovi tega, kar smo z grafov lahko razbrali, lahko rečemo, da se predznaki koeficientov SNV-jev pri uporabi linearne regresije ohranjajo.

4.5.2 Logistična regresija

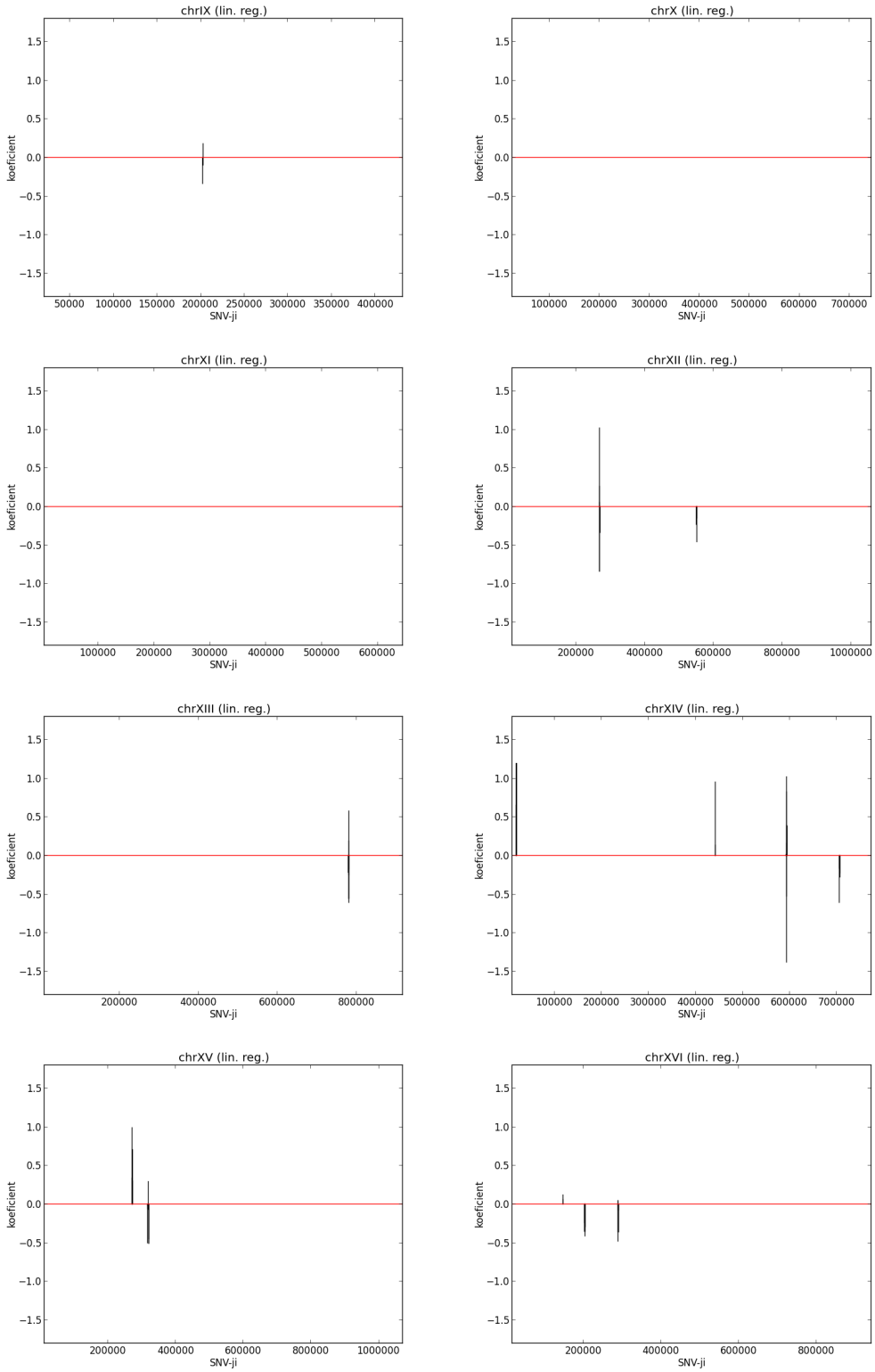
Oglejmo si še grafe, kjer so izrisani koeficienti, dobljeni z uporabo logistične regresije. Ta za vsak vzorec izračuna pripadajoče koeficiente SNV-jev. Kot smo že napisali, nas zanimajo le koeficienti vzorcev s skrajnima fenotipoma ($F7JP\_01$ in IP). To pomeni, da moramo pri logistični regresiji izrisati dva grafa na kromosom. Na zgornjem grafu prikažemo koeficiente vzorca $F7JP\_01$, na spodnjem pa koeficiente vzorca IP.

Primerjajmo zdaj pare grafov (slike 4.9, 4.10 in 4.11) s pari grafov (slike 4.6, 4.7 in 4.8). Ker je opazovana združitev skupin genov sestavljena iz 93 genov (485 SNV-jev), lahko opravimo več primerjav. Koeficienti SNV-jev obdržijo predznak ne glede na to, katera para grafov primerjamo.

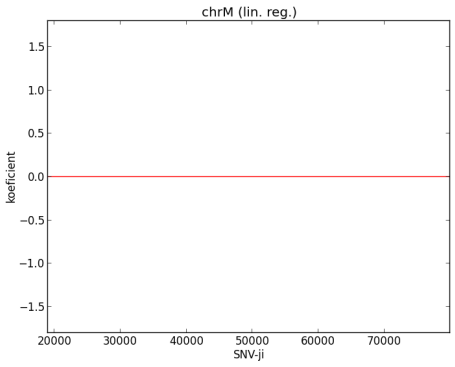
Po primerjavi vseh grafov koeficientov SNV-jev in analizi primerjav lahko potegnemo že kar gotove zaključke. Napovedni model že na veliki bazi atributov (vsi SNV-ji) določi predznake njihovih koeficientov.



Slika 4.6: Vrednosti koeficientov, ko uporabimo SNV-je najboljše združitve skupin in linearno regresijo (chrI-chrVIII).



Slika 4.7: Vrednosti koeficientov, ko uporabimo SNV-je najboljše združitve skupin in linearno regresijo (chrIX-chrXVI).



Slika 4.8: Vrednosti koeficientov, ko uporabimo SNV-je najboljše združitve skupin in linearno regresijo (chrM).

4.6 Točnost rangiranja

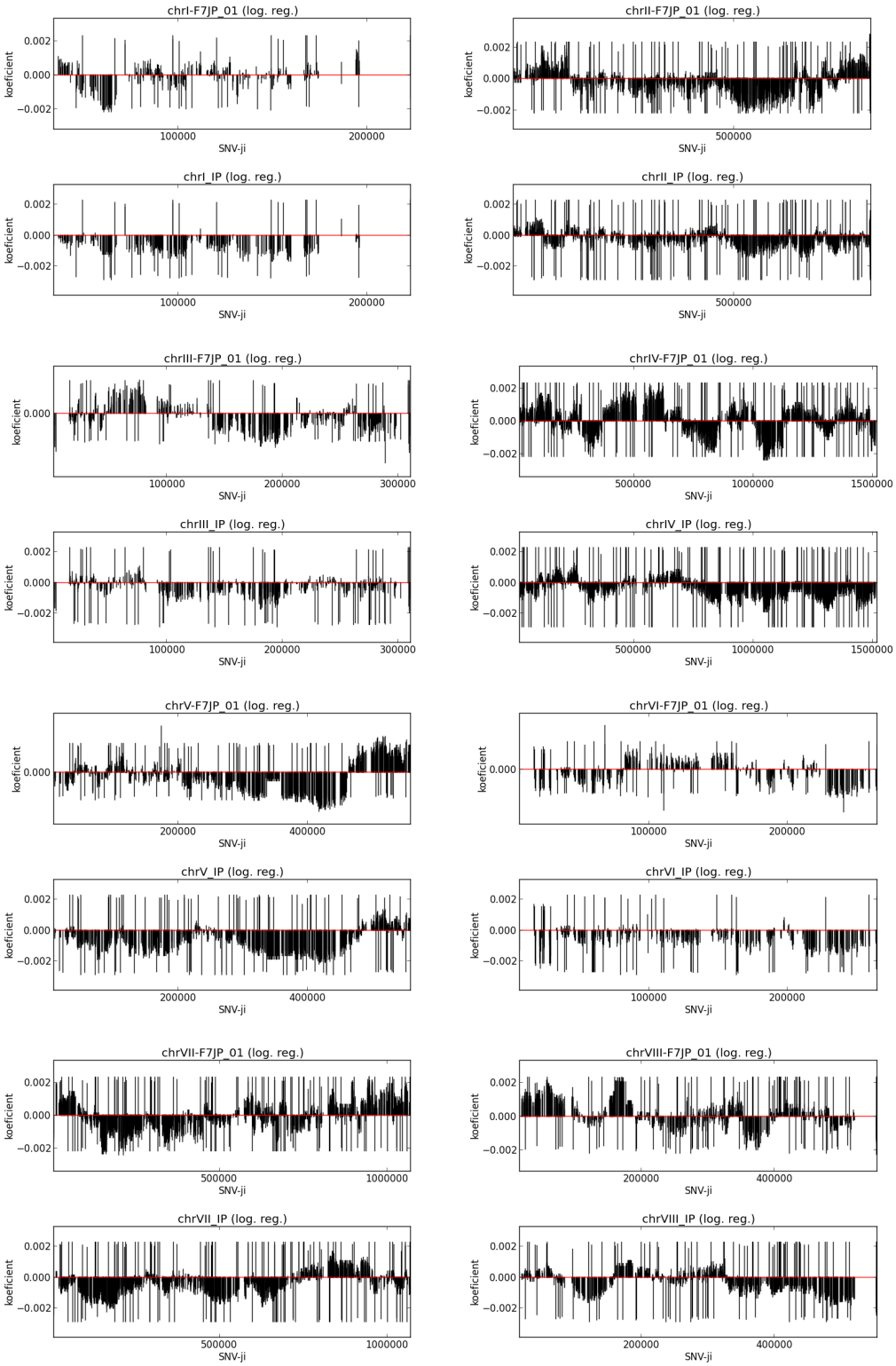
Do zdaj smo se ukvarjali le s klasificiranjem vzorcev (napovedovanjem njihovih fenotipov). V tem poglavju pa smo preučili uspešnost rangiranja vzorcev. Fenotipe vzorcev smo najprej diskretizirali v nekaj razredov. Pokazali smo, koliko vzorcev potrebujemo za grajenje uspešnega napovednega modela in na kakšen način jih moramo izbirati.

Za rangiranje vzorcev smo uporabljali izključno logistično regresijo. Ta napovedni model za vsak primer izračuna, kolikšne so verjetnosti pripadnosti v vsak možen razred. Vzorec potem napovemo v tisti razred, za katerega je verjetnost pripadnosti največja. Glavni razlog za uporabo logistične regresije je, da lahko s pomočjo teh verjetnosti vzorce rangiramo znotraj istega razreda.

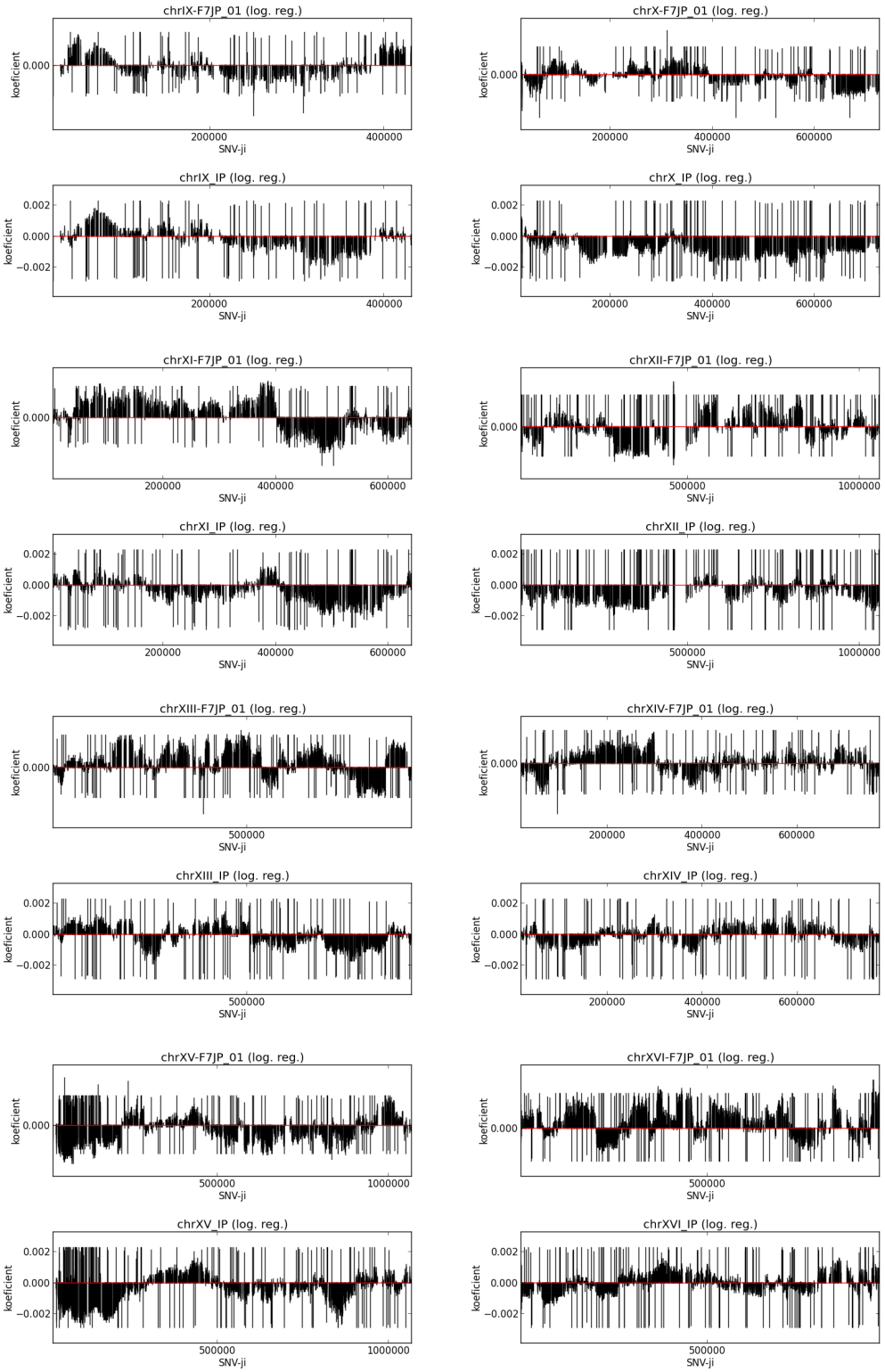
4.6.1 Točnost rangiranja diskretiziranih fenotipov

Preverimo, kako se točnost rangiranja vzorcev spreminja, ko fenotipe vzorcev diskretiziramo na več različnih števil razredov (na 2, 3, 5 in 7 razredov). Postopek, ki ga za to uporabimo, je naslednji:

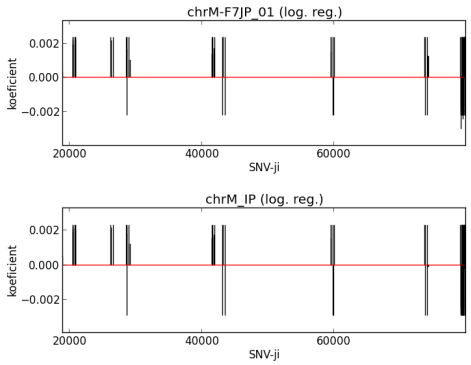
1. Določimo število diskretiziranih razredov k in fenotipe vzorcev spremenimo tako, da bodo njihove vrednosti na intervalu $[1, ..., k]$.



Slika 4.9: Vrednosti koeficientov, ko uporabimo vse SNV-je in logistično regresijo (chrI-chrVIII).



Slika 4.10: Vrednosti koeficientov, ko uporabimo vse SNV-je in logistično regresijo (chrIX-chrXVI).



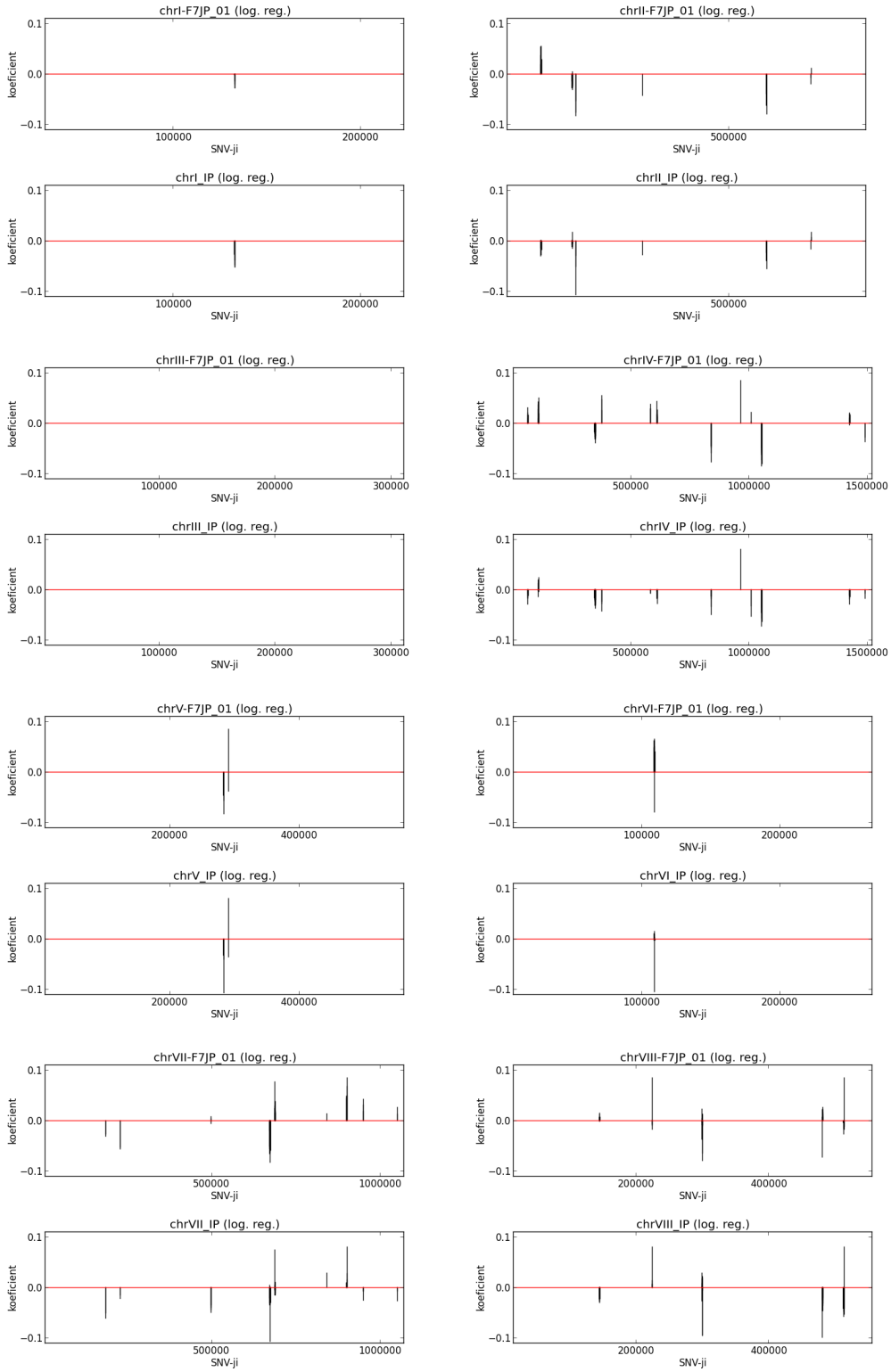
Slika 4.11: Vrednosti koeficientov, ko uporabimo vse SNV-je in logistično regresijo (chrM).

2. Vzamemo eno izmed 100 najboljših združitev skupin genov, ki smo jih dobili z uporabo logistične regresije.

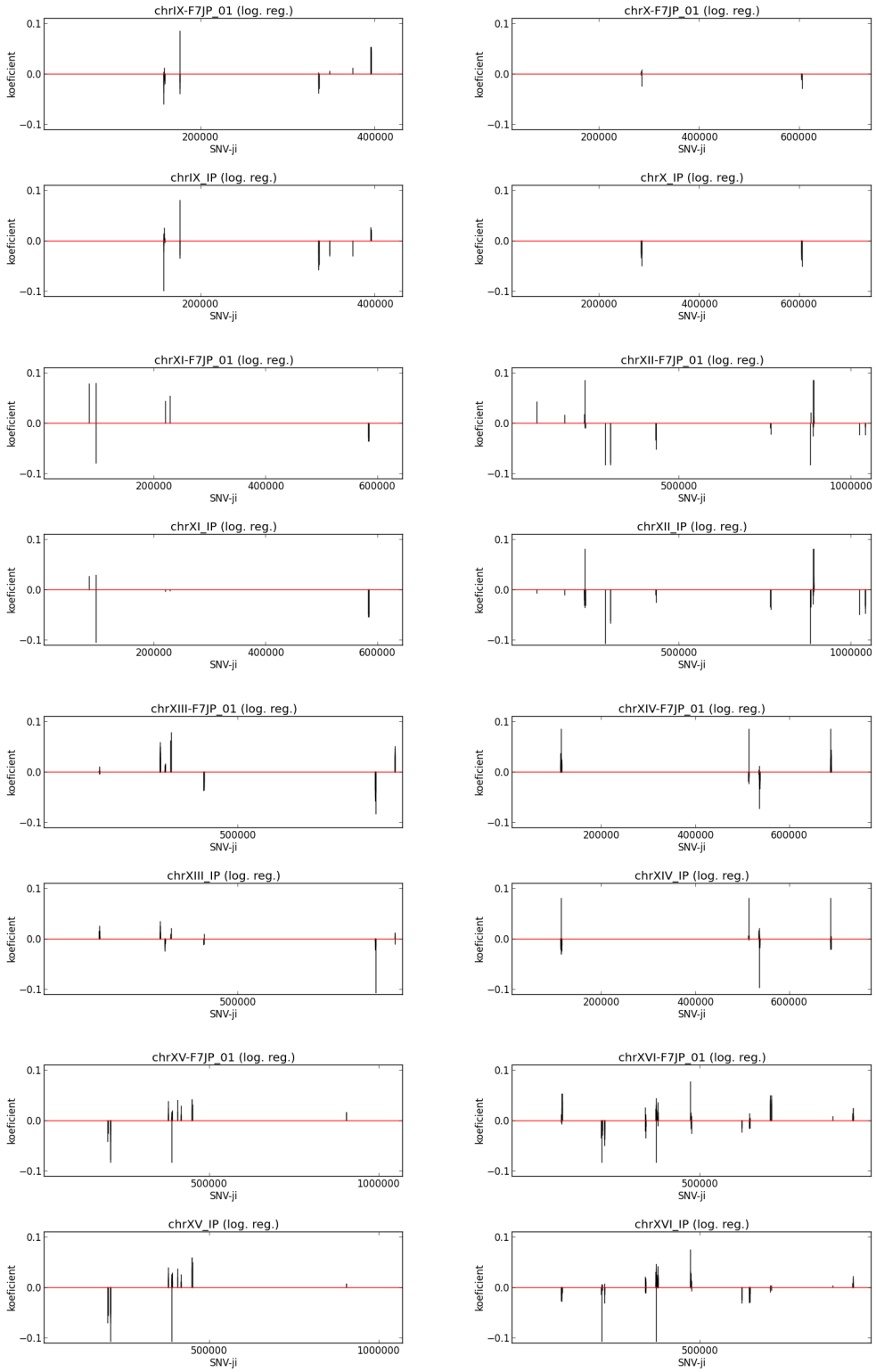
3. Na izbrani združitvi opravimo 26-kratno prečno preverjanje (definicija 3.7) in vzporedno z logistično regresijo napovedujemo, v kateri razred spada vsak vzorec s_i. To pomeni, da vzorec, ki mu je napovedan fenotip F_{s_i}, dodamo v razred $cl_{F_{s_i}}$. Ker bomo kasneje potrebovali pozicijo vsakega vzorca $idx(s_i)$ v prvotnem seznamu, si jih v tem koraku zapomnimo.

4. Ker je v vsakem razredu več vzorcev, jih moramo rangirati znotraj vsakega razreda. To naredimo tako, da si za vsak vzorec zapomnimo verjetnost pripadnosti v razred: $P(s_i \in cl_{F_{s_i}})$. Vzorce potem uredimo na tak način:

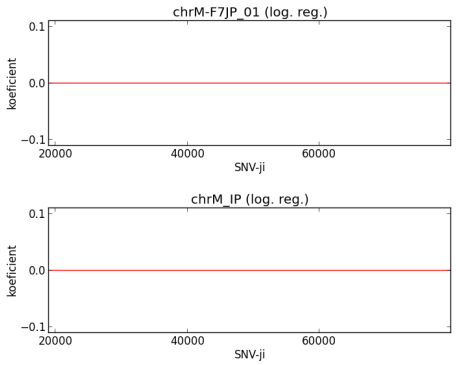
   - Vzorce v razredih cl_j, kjer je $j = F_{s_i}$ za vsak $s_i \in cl_j$, za katere velja
   $$j \leq \lceil \frac{max(cles)}{2} \rceil$$
   sortiramo **padajoče**. To počnemo, ker večja verjetnost pripadnosti razredu pomeni, da sta fenotip in zato tudi rang vzorca bolj



Slika 4.12: Vrednosti koeficientov, ko uporabimo SNV-je najboljše združitve skupin in logistično regresijo (chrI-chrVIII).



Slika 4.13: Vrednosti koeficientov, ko uporabimo SNV-je najboljše združitve skupin in logistično regresijo (chrIX-chrXVI).



Slika 4.14: Vrednosti koeficientov, ko uporabimo SNV-je najboljše združitve skupin in logistično regresijo (chrM).

skrajna, torej v tem primeru nižja. Ureditev vzorcev znotraj razreda je torej takšna:

$$\begin{cases} rang(s_i) \leq rang(s_l), & P(F_{s_i} \in cl_j) \geq P(F_{s_l} \in cl_j) \\ rang(s_i) > rang(s_l), & \text{sicer.} \end{cases}$$

- Vzorce v preostalih razredih sortiramo **naraščajoče**, ker so zdaj višji fenotipi in rangi vzorcev bolj skrajni. Za vsak vzorec s_i znotraj teh razredov torej velja:

$$\begin{cases} rang(s_i) \leq rang(s_l), & P(F_{s_i} \in cl_j) \leq P(F_{s_j} \in cl_j) \\ rang(s_i) > rang(s_l), & \text{sicer.} \end{cases}$$

Ko določamo rang vsakega vzorca v razredu cl_j, moramo upoštevati, da velja:

$$rang(s_i \in cl_j) \in [1 + max(rang(cl_{j-1})), ..., len(cl_j) + max(rang(cl_{j-1}))]\,;$$

kjer je $max(rang(cl_0)) = 0$.

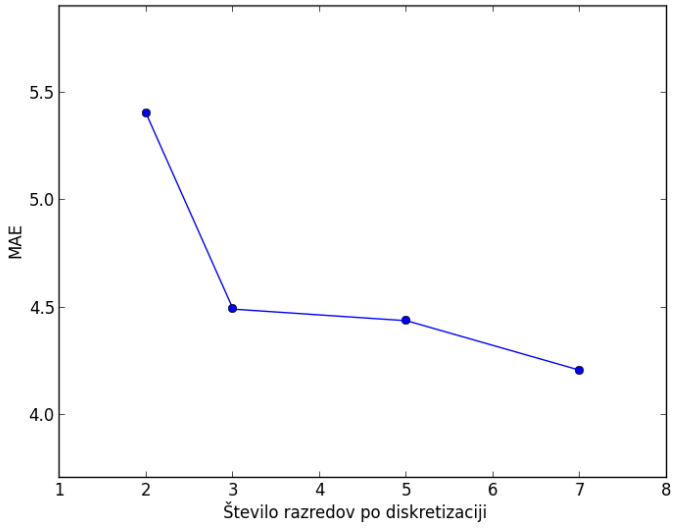
5. Zdaj razrede cl_j združimo tako, da rang vsakega vzorca postavimo na pozicijo, kjer je ta prvotno bil:

$$List\,[idx(s_i)] = rang(s_i).$$

6. Ostane nam le še preverjanje točnosti rangiranja, ki jo preverimo z napakama MAE (definicija 3.9) in MSE (definicija 3.8).

7. Korake od 2.-6. ponovimo za vsako izmed 100 najboljših združitev skupin genov, dobljenih z uporabo logistične regresije.

Postopek smo pognali za različne k-je (števila diskretiziranih razredov). Opazovali smo predvsem odvisnost napak MAE in MSE od števila diskretiziranih razredov. Zaradi tega so izrisane vrednosti napak MAE na garfu (slika 4.15) določene kot povprečje najboljših deset napak MAE, dobljenih po tem postopku. Z grafa je razvidno, da je napaka MAE obratno sorazmerna s številom razredov po diskretizaciji.



Slika 4.15: Graf, ki prikazuje odvisnost napake MAE od števila razredov po diskretizaciji.

4.6.2 Izbor vzorcev za uspešno rangiranje

V tem razdelku smo se ukvarjali s problemom, kako izbirati vzorce in koliko jih potrebujemo za izgradnjo uspešnega napovednega modela. Cilj je izbrati

čim manj vzorcev. Primerjali smo dva načina izbora vzorcev. Pri prvem napovedni model zgradimo tako, da iz vsakega diskretiziranega razreda naključno izberemo en vzorec. Pri drugem pa za gradnjo napovednega modela vzamemo vse vzorce iz obeh ekstremnih razredov (z najmanjšim in največjim fenotipom). Najprej si oglejmo postopek, ki za grajenje napovednega modela uporabi prvi način izbiranja vzorcev (število vseh vzorcev je n):

1. Določimo število diskretiziranih razredov k in fenotipe vzorcev spremenimo tako, da bodo njihove vrednosti na intervalu $[1, ..., k]$.

2. Naključno izberemo po en vzorec s_i iz vsakega razreda cl_j. Če je število razredov k, potem izberemo ravno toliko vzorcev. Te vzorce uporabimo za gradnjo napovednega modela.

3. Ostane nam še $n - k$ vzorcev, ki jih hočemo rangirati, napovedane range pa nato primerjati z dejanskimi. Zaradi tega moramo seznam dejanskih rangov spremeniti tako, da v njem ostanejo le vzorci, katerih range bomo napovedovali. Nato moramo tem vzorcem popraviti range tako, da jih rangiramo na intervalu $[1, ..., n - k]$. To naredimo tako, da:

   (a) s seznama s_1 s prvotnimi rangi izbrišemo vse vzorce, izbrane v prejšnjem koraku, in dobimo seznam s_2,

   (b) inicializiramo ničelni seznam s_3, dolžine $n - k$, in $rang = 1$,

   (c) poiščemo indeks $idx(min(s_2))$ minimalnega elementa v seznamu s_2 in naredimo: $s_3\,[idx(min(s_2))] = rang$,

   (d) minimalni element v s_2 spremenimo v večjega od vseh: $s_2\,[idx(min(s_2))] = 100$ in povečamo rang: $rang+ = 1$,

   (e) koraka (c) in (d) ponavljamo, dokler je $rang < len(s_2)$.

4. Vzamemo eno izmed 100 najboljših združitev skupin genov, ki smo jih dobili pri napovedovanju fenotipov za vse vzorce z logistično regresijo.

5. Z naključno izbranimi k vzorci in z izbrano združitvijo skupin genov zgradimo napovedni model. Z uporabo logistične regresije za vsakega izmed ostalih $n - k$ vzorcev napovemo, v kateri razred spada. Obenem si zapomnimo tudi, kakšna je verjetnost pripadnosti vzorca s_i v napovedani razred cl_j: $P(s_i \in cl_j)$.

6. Preden lahko preverimo pravilnost rangiranja vzorcev, moramo najprej rangirati vzorce znotraj vsakega razreda. To naredimo tako, da za vsak vzorec v razredu primerjamo verjetnosti pripadnosti v ta razred.

   (a) Vzorce v razredih cl_j, za katere velja:

   $$j \leq \lceil \frac{max(cles)}{2} \rceil$$

   sortiramo padajoče tako, da velja

   $$\begin{cases} rang(s_i) \leq rang(s_l), & P(F_{s_i} \in cl_j) \geq P(F_{s_l} \in cl_j) \\ rang(s_i) > rang(s_l), & \text{sicer.} \end{cases}$$

   (b) Vzorce v preostalih razredih sortiramo naraščajoče tako, da velja

   $$\begin{cases} rang(s_i) \leq rang(s_l), & P(F_{s_i} \in cl_j) \leq P(F_{s_j} \in cl_j) \\ rang(s_i) > rang(s_l), & \text{sicer.} \end{cases}$$

   Ko določamo rang vsakega vzorca v razredu cl_j, moramo upoštevati, da velja:

   $$rang(s_i \in cl_j) \in [1 + max(rang(cl_{j-1})), ..., len(cl_j) + max(rang(cl_{j-1}))]\,;$$

   kjer je $max(rang(cl_0)) = 0$.

7. Vse razrede cl_j združimo tako, da rang vsakega vzorca postavimo na pozicijo, na kateri je ta prvotno bil:

   $$List\,[idx(s_i)] = rang(s_i).$$

8. Zdaj lahko preverimo natančnost napovedanega rangiranja. Izmerimo jo z napakama MAE in MSE.

9. Korake od 4. naprej ponovimo za vsako izmed 100 najboljših združitev skupin.

10. Ker vzorce izbiramo naključno, postopek od 2. koraka naprej ponovimo večkrat (250-krat), da dobimo povprečno točnost rangiranja.

Opišimo še postopek napovedovanja rangov, če za grajenje napovednih modelov uporabimo vse vzorce iz ekstremnih razredov (z najmanjšim in največjim fenotipom).

1. Določimo število diskretiziranih razredov k in fenotipe vzorcev spremenimo tako, da so njihove vrednosti na intervalu $[1, ..., k]$.

2. Za napovedni model vzamemo vse vzorce iz razredov $cl_{min(j)}$ in $cl_{max(j)}$. Recimo, da je takih vzorcev l.

3. Vzamemo eno izmed 100 najboljših združitev skupin genov, ki smo jih dobili pri napovedovanju fenotipov za vse vzorce z logistično regresijo.

4. Z izbranimi vzorci in izbrano združitvijo skupin genov izgradimo napovedni model. Z uporabo logistične regresije za vsakega izmed ostalih $n - l$ vzorcev napovemo, v kateri razred spada. Obenem si zapomnimo tudi, kakšna je verjetnost pripadnosti vzorca s_i v napovedani razred cl_j: $P(s_i \in cl_j)$.

5. Kot pri prejšnjem postopku moramo tudi tu najprej rangirati vzorce znotraj vsakega razreda. To naredimo tako, da za vsak vzorec v razredu primerjamo verjetnosti pripadnosti v ta razred.

   (a) Vzorce v razredu $cl_{min(j)}$ sortiramo **padajoče** tako, da velja:

$$\begin{cases} rang(s_i) \leq rang(s_l), & P(F_{s_i} \in cl_j) \geq P(F_{s_l} \in cl_j) \\ rang(s_i) > rang(s_l), & \text{sicer.} \end{cases}$$

(b) Vzorce v razredu $cl_{max(j)}$ sortiramo **naraščajoče** tako, da velja:

$$\begin{cases} rang(s_i) \leq rang(s_l), & P(F_{s_i} \in cl_j) \leq P(F_{s_j} \in cl_j) \\ rang(s_i) > rang(s_l), & \text{sicer.} \end{cases}$$

Ko določamo rang vsakega vzorca v obeh razredih, moramo upoštevati, da velja:

$$rang(s_i \in cl_{min(j)}) \in \left[1, ..., len(cl_{min(j)}))\right] \text{ in}$$

$$rang(s_i \in cl_{max(j)}) \in \left[1 + len(cl_{min(j)}), ..., len(cl_{max(j)})) + len(cl_{min(j)})\right].$$

6. Zdaj lahko preverimo natančnost napovedanega rangiranja. Izmerimo jo z napakama MAE in MSE, kot smo tega že vajeni.

7. Korake od 3. do 6. ponovimo za vsako izmed 100 najboljših združitev skupin.

Primerjajmo najprej rezultate, ko fenotipe vzorcev diskretiziramo v pet razredov (tabela 4.18). Ker nas je zanimalo, kateri postopek je povprečno boljši, smo primerjali povprečne napake MAE. Te smo izračunali iz deset najnižjih dobljenih napak MAE. Število vzorcev uporabljeno za grajenje napovednega modela, po drugem postopku (*ekstremni*) je dvakrat večje. Zaradi tega smo se odločili prvi postopek ponoviti tako, da smo iz vsakega razreda naključno izbrali po dva vzorca. Iz tako dobljenih rezultatov sledi, da je ob istem številu vzorcev boljše, če jih izbiramo naključno iz vsakega razreda.

Vzorci	**Število vzorcev**	**Povprečna MAE**	**Razlika**
naključni	5	5.2031	0.8656
ekstremni	10	4.3375	
naključni	10	3.5575	-0.78
ekstremni	10	4.3375	

Tabela 4.18: Primerjava obeh postopkov izbiranja vzorcev, ko je število diskretiziranih razredov enako 5.

Poglejmo še rezultate, če fenotipe vzorcev diskretiziramo v sedem razredov (tabela 4.19). Spet smo primerjali povprečne napake MAE, ki smo jih izračunali iz deset najnižjih dobljenih napak MAE. Iz prvih dveh vrstic tabele je razvidno, da je že pri naključnem izboru enega vzorca iz vsakega razreda napaka manjša (kljub temu, da je napovedni model zgrajen z enim vzorcem manj). Če pa iz vsakega razreda izberemo po dva naključna vzorca, dosežemo že zelo visoko točnost rangiranja.

Vzorci	**Število vzorcev**	**Povprečna MAE**	**Razlika**
naključni	7	4.5070	0.6486
ekstremni	8	5.1556	
naključni	14	2.3373	2.8183
ekstremni	8	5.1556	

Tabela 4.19: Primerjava obeh postopkov, ko je število diskretiziranih razredov enako 7.

Zaključimo lahko, da je najbolje iz vsakega razreda izbrati vsaj eden vzorec. Velja tudi, da lahko z izbiro naključnih dveh vzorcev iz vsakega razreda ostale vzorce rangiramo precej dobro. Še posebej to velja za primer, ko smo fenotipe vzorcev diskretizirali v 7 razredov.

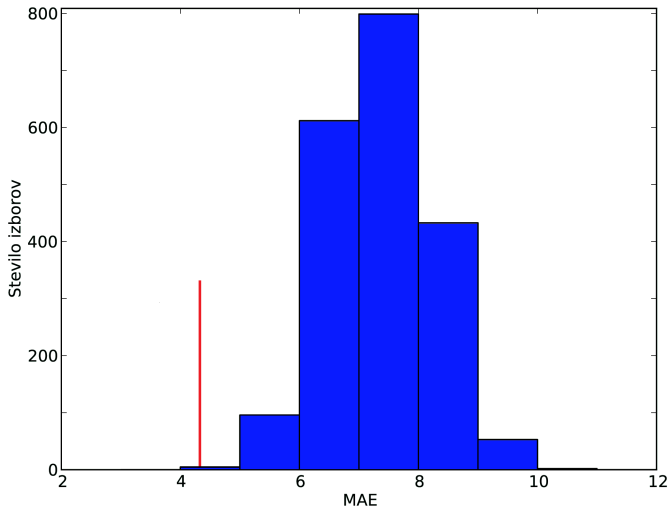
4.7 Posamezniki in celotna populacija

V tem poglavju smo preverili, ali so podatki o manjvrednem staršu (IP), večvrednem staršu (SP) in populaciji $F1\_pool$ dovolj za gradnjo uspešnega napovednega modela, ki bo pravilno rangiral posameznike ($F7JP\_01$, ..., $F7JP\_26$) iz populacije $F7\_pool$. Ugotovili smo že, da za zelo dobro rangiranje potrebujemo vsaj deset vzorcev. Ti morajo biti izbrani tako, da iz vsakega diskretiziranega razreda naključno izberemo vsaj enega.

Torej, imamo le tri vzorce za grajenje napovednega modela. Problem je tako podoben situaciji, ko fenotipe vzorcev diskretiziramo v tri razrede.

Zato smo kot napovedni model uporabili le logistično regresijo. Na podlagi slike (slika 4.15) smo pričakovali, da bodo najboljši rezultati dosegali napako MAE približno 4.5.

Uporabili smo zelo podoben postopek, kot je opisan v podpoglavju 4.5.1, in ga zato ne bomo še enkrat opisali. Edina razlika je, da zdaj fenotipe vzorcev diskretiziramo v razrede 3, 10 in 26 (namesto 1, 2, 3). To naredimo, ker so fenotipi F_i primerov v napovednem modelu: $F_{IP} = 26$, $F_{SP} = 10$ in $F_{F1\_pool} = 3$.



Slika 4.16: Napaka MAE predlaganega postopka je označena z rdečo, navpično črto in znaša 4.3.

Najprej poglejmo rezultate, ko smo gene za grajenje napovednega modela izbirali naključno. Združitve skupin genov, s katerimi smo naključne izbore primerjali, so dolge približno 100 genov. Zato morajo biti tudi naključni izbori sestavljeni iz 100 genov. Pri naključnih postopkih nas po navadi zanima predvsem, kako dobri so v povprečju. Zato smo postopek naključnega izbora genov ponovili 2000-krat. Tako dobljene rezultate smo predstavili s histogramom (slika 4.16), ki pokaže porazdelitev napake MAE.

Primerjajmo rezultate histograma na sliki 4.16 s svojimi rezultati (tabela 4.20). Te smo dobili tako, da smo napovedne modele zgradili z naj-

Postopek	**Rezultat**	**MAE**	**MSE**	**P(MAE)**
predznanje	najboljši	4.30	31.96	0%
naključen	povprečen	7.39	79.10	49.85%
naključen	najboljši	4.57	42.65	0.0005%

Tabela 4.20: Primerjava postopka izbora genov na podlagi predznanja s postopkom naključnega izbora genov za rangiranje posameznikov iz populacije *F7_pool*.

boljšimi združitvami skupin genov, dobljenih v razdelku 4.1.3 z uporabo logistične regresije. Vrednosti v stolpcu **P(MAE)** pomenijo verjetnost, da z naključnim postopkom dobimo manjšo ali enako napako MAE.

Kot smo napisali na začetku poglavja, je napaka MAE, dobljena z našim postopkom, približno 4.5, kar kaže na zmožnost gradnje sorazmerno uspešnega napovednega modela na podlagi le nekaj vzorcev. Iz teh vrednosti napak MAE lahko sklepamo, da z uporabo predznanja tudi v tem primeru dosegamo boljše rezultate kot z naključnim izborom genov.

Poglavje 5

Sklepne ugotovitve

Na začetku diplomske naloge smo si zadali veliko ciljev. Poiskali smo čim manjši nabor genov in SNV-jev, s katerimi kar se da dobro klasificiramo dane vzorce. Poleg tega smo ugotovili, kako izbirati vzorce in koliko jih moramo izbrati za izgradnjo uspešnega modela za napovedovanje fenotipa. Za konec smo prikazali še natančnost rangiranja, ko napovedni model zgradimo le na podlagi podatkov o začetnih starših in prvi generaciji (vzorci IP, SP in $F1\_pool$).

Eden izmed ciljev je bil tudi ugotoviti, kateri geni in SNV-ji so najbolj povezani s fenotipom vzorcev in od katerega od staršev morajo izvirati, da bo fenotip vzorca dober. Za oba napovedna modela (linearno in logistično regresijo) smo pokazali načina, s katerima lahko to ugotovimo. Poleg tega smo preverili tudi, kako izbira naborov SNV-je vpliva na napovedno točnost zgrajenih modelov.

Za doseganje vseh rezultatov smo uporabili predznanje v podatkovnih zbirkah GO in KEGG. Gene smo lahko razvrstili v skupine tako, da smo v isto skupino dali gene, ki sodelujejo pri istih procesih. Dokazati smo želeli, da tak način izbiranja genov privede do dobrih napovednih modelov.

Diplomsko delo je bilo precej obsežno in v njem smo se srečali s številnimi empiričnimi vprašanji.

5.1 Klasificiranje in rangiranje posameznikov in populacij

Iz rezultatov je razvidno, da se pri klasificiranju vzorcev veliko bolje izkaže linearna regresija. Z njo smo dosegli že zavidljivo stopnjo natančnosti (vrednost napake MAE je bila krepko manjša od 1.0, pri razponu zveznega razreda 1..29). Poleg tega so združitve skupin genov, dobljene s tem napovednim modelom, sestavljene iz manjšega števila genov. Tudi klasificiranje vzorcev z logistično regresijo je bilo uspešno.

Za rangiranje vzorcev smo kot napovedni model uporabili le logistično regresijo. Vzorce smo rangirali tako, da smo njihove fenotipe najprej diskretizirali na nekaj razredov. Ugotovili smo, da uspešnost rangiranja vzorcev raste z večanjem števila diskretiziranih razredov. Pokazali smo tudi, da moramo vzorce izbirati tako, da iz vsakega diskretiziranega razreda vzamemo vsaj enega in tako zagotovimo reprezentativen vzorec.

S temi rezultati smo lahko zanesljivo ocenili točnost rangiranja posameznikov, če napovedni model zgradimo samo na podlagi vzorcev IP, SP in $F1\_pool$. Vrednost napake MAE je bila manjša od 4.5.

5.2 Pomembnost genov in SNV-jev

V dobljenih združitvah skupin genov, ki so najboljše napovedovale fenotipe vzorcev, smo poiskali najbolj pomembne gene in SNV-je. To smo naredili s pregledovanjem koeficientov, ki jih vrne napovedni model (linearna ali logistična regresija). Izkazalo se je, da najbolj pomembni SNV-ji v večini tudi pripadajo najbolj pomembnim genom.

S pomočjo koeficientov smo lahko določili, od katerega starša mora biti podedovan posamezni gen ali SNV, da je fenotip vzorca dober. Pokazali smo, da mora biti večina genov in SNV-jev podedovanih od večvrednega starša (SP). To je tudi logično, saj je njegov fenotip veliko boljši od fenotipa manjvrednega starša. Po drugi strani pa je treba podedovati tudi nekaj genov

in SNV-jev slabšega starša (IP), če želimo doseči še boljši fenotip.

Za koeficiente SNV-jev nas je zanimalo ali se njihovi predznaki spreminjajo z izbiro napovednega modela. Primerjali smo koeficiente modelov, ko smo za gradnjo napovednega modela vzeli vse SNV-je, s koeficienti, ko smo za gradnjo napovednega modela vzeli le SNV-je iz najboljše združitve skupin genov. S precejšnjo gotovostjo lahko trdimo, da se predznaki koeficientov ne spremenijo. To pomeni, da je že v začetnem modelu v veliki meri določeno, od katerega starša (IP ali SP) mora biti podedovan posamezen SNV.

5.3 Uporabnost predznanja

Uporaba predznanja iz zbirk GO in KEGG se je izkazala za zelo uspešno. Z uporabo funkcijskih pripisov genov smo lahko v posamezne skupine razvrstili gene, ki sodelujejo pri istih procesih. Kakovost izbire in uporabe opisanega predznanja smo ovrednotili s številnimi referenčnimi vrednostmi (naključni izbor genov, povprečni vektorji skupin, itd). Pri vsaki smo z uporabo predznanja dosegli bistveno boljše rezultate.

Za najboljši združitvi skupin genov smo pokazali tudi, kateri izraz GO ju najbolje opisuje. Ker imamo tri različne tipe opisov GO (biološki procesi, predeli celice, molekularne funkcije), smo se omejili na izpis le najboljšega za vsak tip. Najboljši so bili sestavljeni iz nabora genov, za katerega je verjetnost, da ga dobimo z naključnim izborom, zelo majhna. To smo ocenili s p-vrednostjo in deležem napačno pozitivnih odkritij (FDR).

5.4 Nadaljnje delo

Kot smo omenili, je bilo treba v diplomski nalogi sprejeti mnogo empiričnih odločitev. Še bolj poglobljeno iskanje optimalnih parametrov bi bilo časovno in računsko zelo zahtevno, hkrati bi povečali nevarnost pretiranega prilegaanja podatkom. Vsekakor dopuščamo možnost, da obstaja še kakšna dobra, neodkrita rešitev.

Rezultati, dobljeni z drugačnima napovednima modeloma, so se precej razlikovali, saj so najboljše združitve skupin genov bile sestavljene iz drugačnih genov. Zato obstaja možnost, da bi z drugim napovednim modelom odkrili drugačne nabore genov, ki tudi dobro napovedujejo fenotip. Natančnost klasificiranja vzorcev, ki smo jo dobili z linearno regresijo, bi v vsakem primeru težko presegli.

Kakovost opisanih metod in rezultatov bi bilo vredno oceniti še na drugih podobnih podatkih.

Literatura

[1] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

[2] J. A. Blake, M. Dolan, H. Drabkin, D. P. Hill, Ni Li, D. Sitnikov, S. Bridges, S. Burgess, T. Buza, F. McCarthy, D. Peddinti, L. Pillai, S. Carbon, H. Dietze, A. Ireland, S. E. Lewis, C. J. Mungall, P. Gaudet, R. L. Chrisholm, P. Fey, W. A. Kibbe, S. Basu, D. A. Siegele, B. K. McIntosh, D. P. Renfro, A. E. Zweifel, J. C. Hu, N. H. Brown, S. Tweedie, Y. Alam-Faruque, R. Apweiler, A. Auchinchloss, K. Axelsen, B. Bely, M. C Blatter, C. Bonilla, L. Bouguerleret, E. Boutet, L. Breuza, A. Bridge, W. M. Chan, G. Chavali, E. Coudert, E. Dimmer, A. Estreicher, L. Famiglietti, M. Feuermann, A. Gos, N. Gruaz-Gumowski, R. Hieta, C. Hinz, C. Hulo, R. Huntley, J. James, F. Jungo, G. Keller, K. Laiho, D. Legge, P. Lemercier, D. Lieberherr, M. Magrane, M. J. Martin, P. Masson, P. Mutowo-Muellenet, C. O'Donovan, I. Pedruzzi, K. Pichler, D. Poggioli, P. Porras Millán, S. Poux, C. Rivoire, B. Roechert, T. Sawford, M. Schneider, A. Stutz, S. Sundaram, M. Tognolli, I. Xenarios, R. Foulgar, J. Lomax, P. Roncaglia, V. K. Khodiyar, R. C. Lovering, P. J. Talmud, M. Chibucos, M. Gwinn Giglio, H. Y Chang, S. Hunter, C. McAnulla, A. Mitchell, A. Sangrador, R. Stephan, M. A. Harris, S. G. Oliver, K. Rutherford, V. Wood, J. Bahler, A. Lock, P. J. Kersey, D. M. McDowall, D. M. Staines, M. Dwinell, M. Shimoyama, S. Laulederkind, T. Hayman, S. J Wang, V. Petri, T. Lowry, P. D'Eustachio, L. Matthews, R. Balakri-

shnan, G. Binkley, J. M. Cherry, M. C. Costanzo, S. S. Dwight, S. R. Engel, D. G. Fisk, B. C. Hitz, E. L. Hong, K. Karra, S. R. Miyasato, R. S. Nash, J. Park, M. S. Skrzypek, S. Weng, E. D. Wong, T. Z. Berardini, E. Huala, H. Mi, P. D. Thomas, J. Chan, R. Kishore, P. Sternberg, K. Van Auken, D. Howe, and M. Westerfield. Gene ontology annotations and resources. *Nucleic Acids Res*, 41(Database issue):D530–5, 1 2013.

[3] T. Curk, J. Demsar, Q. Xu, G. Leban, U. Petrovic, I. Bratko, Shaulsky G., and B. Zupan. Microarray data mining with visual programming.

[4] Jorge Duitama, Juan Camilo Quintero, Daniel Felipe Cruz, Constanza Quintero, Georg Hubmann, Maria R. Foulquié-Moreno, Kevin J. Verstrepen, Johan M. Thevelein, and Joe Tohme. An integrated framework for discovery and genotyping of genomic variants from high-throughput sequencing experiments. *Nucleic Acids Res*, 42(6):e44, 4 2014.

[5] Jorge Duitama, Aminael Sánchez-Rodríguez, Annelies Goovaerts, Sergio Pulido-Tamayo, Georg Hubmann, María R. Foulquié-Moreno, Johan M. Thevelein, Kevin J. Verstrepen, and Kathleen Marchal. Improved linkage analysis of quantitative trait loci using bulk segregants unveils a novel determinant of high ethanol tolerance in yeast. *BMC Genomics*, 15(1):207, 2014.

[6] Christopher R. Genovese. A tutorial on false discovery control. http://www.stat.cmu.edu/ genovese/talks/hannover1-04.pdf.

[7] P. M. Hooper. What is a p-value? http://www.stat.ualberta.ca/ hooper/teaching/misc/Pvalue.pdf.

[8] D. W. Hosmer and S. Lemeshow. *Applied logistic regression.* John Wiley & Sons, Inc., second edition, 2000.

[9] R. J. Hyndman and A. B. Koehler. Another look at measures of forecast accuracy, 2005.

[10] Minoru Kanehisa, Susumu Goto, Miho Furumichi, Mao Tanabe, and Mika Hirakawa. Kegg for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res*, 38(Database issue):D355–60, 1 2010.

[11] I. Kononenko. *Strojno ucenje*. Zalozba FE in FRI, 2005.

[12] I. Kononenko and M. Kukar. *Machine learning and data mining*. Horwood publishing Limited, 2007.

[13] F. Murtagh and P. Legendre. Ward's hierarchical clustering method: Clustering criterion and agglomerative algorithm. http://arxiv.org/abs/1111.6285v2, 2011.

[14] NIST. Nist/sematech e-handbook of statistical methods. http://www.itl.nist.gov/div898/handbook/, 2013.

[15] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[16] K. Raza. Application of data mining in bioinformatics. *Indian Journal of Computer Science and Engineering*, 1(2):114–118, 2010.

[17] J. Schneider. Cross validation. http://www.cs.cmu.edu/ schneide/tut5/node42.html, 1997.

[18] Steve Swinnen, Kristien Schaerlaekens, Thiago Pais, Jürgen Claesen, Georg Hubmann, Yudi Yang, Mekonnen Demeke, María R. Foulquié-Moreno, Annelies Goovaerts, Kris Souvereyns, Lieven Clement, Françoise Dumortier, and Johan M. Thevelein. Identification of novel causative genes determining the complex trait of high ethanol tolerance in yeast using pooled-segregant whole-genome sequence analysis. *Genome Res*, 22(5):975–84, 5 2012.

[19] Pang-Ning . N. Tan, Michael Steinbach, and Vipin Kumar. *Introduction to data mining*. Pearson Addison Wesley, Boston, 2006.

[20] Wikipedia. Mean absolute error. http://en.wikipedia.org/wiki/Mean_absolute_error, 2013.

[21] Wikipedia. Cluster analysis. http://en.wikipedia.org/wiki/Cluster_analysis, 2014.

[22] Wikipedia. Cross-validation (statistics). http://en.wikipedia.org/wiki/Cross-validation_(statistics), 2014.

[23] Wikipedia. Mean squared error. http://en.wikipedia.org/wiki/Mean_squared_error, 2014.

[24] Wikipedia. Outlier. http://en.wikipedia.org/wiki/Outlier, 2014.

[25] Wikipedia. p-value. http://en.wikipedia.org/wiki/P-value, 2014.

[26] Wikipedia. Q-q plot. http://en.wikipedia.org/wiki/Q-Q_plot, 2014.

[27] Wikipedia. Standard score. http://en.wikipedia.org/wiki/Standard_score, 2014.

[28] Wikipedia. Ward's method. http://en.wikipedia.org/wiki/Ward's_method, 2014.

[29] D. Yekutieli and Y. Benjamini. Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *Journal of Statistical Planning and Inference*, 82(1-2):171–196, 1999.